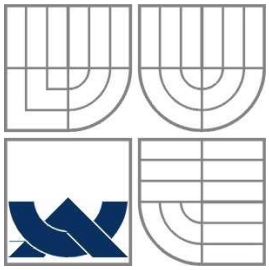
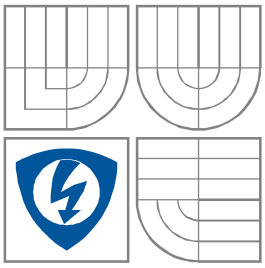


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

SYSTÉM PRO ZPRACOVANÍ SKÓRE Z METOD IDENTIFIKACE PROTEINŮ V TANDEMOVÉ HMOTNOSTNÍ SPEKTROMETRII

SCORING PROCESSING SYSTEM FOR PROTEIN IDENTIFICATION IN TANDEM MASS SPECTROMETRY

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Martin Valla
AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. Ivo Provazník, Ph.D.
SUPERVISOR

BRNO, 2008

Zadání.

LICENČNÍ SMLOUVA
POSKYTOVANÁ K VÝKONU PRÁVA UŽÍT ŠKOLNÍ DÍLO

uzavřená mezi smluvními stranami:

1. Pan/paní

Jméno a příjmení: Martin Valla
Bytem: U Pošty 9, Brno, 625 00
Narozen/a (datum a místo): 22. listopadu 1981 v Brně

(dále jen „autor“)

a

2. Vysoké učení technické v Brně

Fakulta elektrotechniky a komunikačních technologií
se sídlem Údolní 53, Brno, 602 00
jejímž jménem jedná na základě písemného pověření děkanem fakulty:
prof. Ing. Jiří Jan,CSc, předseda rady oboru Biomedicínckéa ekologické inženýrství
(dále jen „nabyvatel“)

Čl. 1

Specifikace školního díla

1. Předmětem této smlouvy je vysokoškolská kvalifikační práce (VŠKP):

   ☐ disertační práce
   ☒ diplomová práce
   ☐ bakalářská práce
   ☐ jiná práce, jejíž druh je specifikován jako .....
   (dále jen VŠKP nebo dílo)

   Název VŠKP: Systém pro zpracovaní skóre z metod identifikace proteinů v tandemové hmotnostní spektrometrii

   Vedoucí/ školitel VŠKP: prof. Ing. Ivo Provazník, Ph.D.
   Ústav: Ústav biomedicínckého inženýrství
   Datum obhajoby VŠKP: 9. 6. 2008

   VŠKP odevzdal autor nabyvateli*:

   ☒ v tištěné formě – počet exemplářů: 2
   ☒ v elektronické formě – počet exemplářů: 2

2. Autor prohlašuje, že vytvořil samostatnou vlastní tvůrčí činností dílo shora popsané a specifikované. Autor dále prohlašuje, že při zpracovávání díla se sám nedostal do rozporu s autorským zákonem a předpisy souvisejícími a že je dílo dílem původním.

3. Dílo je chráněno jako dílo dle autorského zákona v platném znění.

4. Autor potvrzuje, že listinná a elektronická verze díla je identická.

* hodící se zaškrtněte

Článek 2

Udělení licenčního oprávnění

1. Autor touto smlouvou poskytuje nabyvateli oprávnění (licenci) k výkonu práva uvedené dílo nevýdělečně užít, archivovat a zpřístupnit ke studijním, výukovým a výzkumným účelům včetně pořizovaní výpisů, opisů a rozmnoženin.

2. Licence je poskytována celosvětově, pro celou dobu trvání autorských a majetkových práv k dílu.

3. Autor souhlasí se zveřejněním díla v databázi přístupné v mezinárodní síti

   ☒ ihned po uzavření této smlouvy
   ☐ 1 rok po uzavření této smlouvy
   ☐ 3 roky po uzavření této smlouvy
   ☐ 5 let po uzavření této smlouvy
   ☐ 10 let po uzavření této smlouvy
   (z důvodu utajení v něm obsažených informací)

4. Nevýdělečné zveřejňování díla nabyvatelem v souladu s ustanovením § 47b zákona č. 111/ 1998 Sb., v platném znění, nevyžaduje licenci a nabyvatel je k němu povinen a oprávněn ze zákona.

Článek 3

Závěrečná ustanovení

1. Smlouva je sepsána ve třech vyhotoveních s platností originálu, přičemž po jednom vyhotovení obdrží autor a nabyvatel, další vyhotovení je vloženo do VŠKP.

2. Vztahy mezi smluvními stranami vzniklé a neupravené touto smlouvou se řídí autorským zákonem, občanským zákoníkem, vysokoškolským zákonem, zákonem o archivnictví, v platném znění a popř. dalšími právními předpisy.

3. Licenční smlouva byla uzavřena na základě svobodné a pravé vůle smluvních stran, s plným porozuměním jejímu textu i důsledkům, nikoliv v tísni a za nápadně nevýhodných podmínek.

4. Licenční smlouva nabývá platnosti a účinnosti dnem jejího podpisu oběma smluvními stranami.

V Brně dne: 30. května 2008

…….. ……

Nabyvatel Autor

Abstrakt

Cílem práce bylo nalezení metody generující jeden určující parametr pro sjednocení výsledků z různých nástrojů identifikace proteinů v tandemové hmotnostní spektrometrii. Data na výstupu z hmotnostního spektrometru jsou zpracována ve dvou na sobě nezávislých nástrojích Mascot a X!Tandem. Výsledky jsou pak navrženou metodou zpracovány a unifikovány jediným parametrem jednoznačně určujícím identifikaci nalezených proteinů. Navržený skórovací parametr se označuje jako Metaskóre a je výstupem metody, která byla implementována v prostředí MATLAB a prakticky ověřena na reálných datech z veřejně přístupné databáze.

Abstract

The goal of my diploma thesis was finding a suitable method for unifying score values from various protein identification search tools in MS/MS mass spectrometry into one single score value. Data coming from the output of mass spectrometer are processed in two independent search tools Mascot and X!Tandem. These were selected especially for their wide usage in proteomic labs. Both results are evaluated through newly designed function and unified by single valued score clearly identifying found proteins. Newly designed scoring value is called Matascore and function producing this score was implemented in MATLAB. Function and its results were successfully tested by real data available in public databases on the Internet.

Klíčová slova

Ohodnocení, Mascot, X!Tandem, Metaskóre, TMS, peptidová mapa, tandemová hmotnostní spektrometrie.

Keywords

Score, Mascot, X!Tandem, Metascore, MS/MS, peptide mass fingerprint, tandem mass spectrometry.

Citace

VALLA, M. *Systém pro zpracovaní skóre z metod identifikace proteinů v tandemové hmotnostní spektrometrii.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2008. 91 s. Vedoucí diplomové práce prof. Ing. Ivo Provazník, Ph.D.

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Systém pro zpracovaní skóre z metod identifikace proteinů v tandemové hmotnostní spektrometrii jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 30. května 2008 .....
podpis autora

Poděkování

Děkuji vedoucímu diplomové práce prof. Ing. Ivo Provazníkovi, Ph.D. a Ing. Martinovi Plchútovi za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne 30. května 2008 .....
podpis autora

OBSAH

1. ÚVOD

V současné době dochází čím dál více k uplatňování biochemie a biomedicíny jako mezioborových věd. Tato práce je zaměřena na spojení technického a humanitního pohledu na biomedicínské inženýrství. Spojením informatiky a biologie se oblast bioinformatiky velmi rozvíjí. V prostředí plném možností nachází uplatnění řada značně prosperujících firem.

K charakterizaci proteomických dat a rozpoznávání jednotlivých proteinů byly pro tuto práci vybrány techniky bioinformatiky v kombinaci s metodami genetického inženýřství. K pochopení této práce je tedy nezbytná znalost základních pojmů a principů jak z biologie, tak techniky. K objasnění slouží teoretický úvod nazvaný Základy genomiky a proteomiky. Hodně je zde čerpáno z české knihy Dr. Fatimy Cvrčkové [8]. Zbytek práce je se věnuje problematice identifikace neznámých aminokyselin a skórování získaných dat.

K získání proteomických dat je více cest, v této práci jsou data získávána z přístroje nazvaný Tandemový hmotnostní spektrometr. Také se nabízí více možností identifikace, v práci je použit princip srovnání se vzorovou databází. Hodnocení identifikovaných proteinů je možné provést také několika způsoby. V práci jsou vybrány z nejpoužívanějších přístupů dva - X!Tandem a Mascot. Výstupy z těchto nástrojů jsou rozsáhlé peptidové listy, které se od sebe liší. Je žádoucí zavést sjednocující prvek, který v sobě obsahuje přednosti obou přístupů.

2. ZÁKLADY GENOMIKY A PROTEOMIKY

2.1 Molekulární biologie

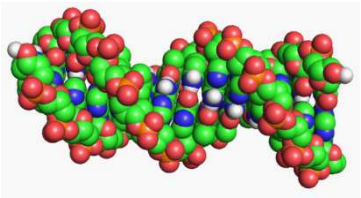
Molekulární biologie je vědní disciplína zabývající se studiem buněčných biologických procesů na jejich molekulární úrovni. Zvláštní pozornost je především věnována studiu funkce makromolekul podílejících se na dědičnosti organizmů, tedy kyseliny deoxyribonukleové (DNA), kyseliny ribonukleové (RNA) a proteinům, jejich vzájemné interakci a regulaci jejich funkce.

Atomová úroveň

Atom je základní částice běžné hmoty, částice, kterou už chemickými prostředky dále nelze dělit a která definuje vlastnosti daného chemického prvku.

Molekulární úroveň

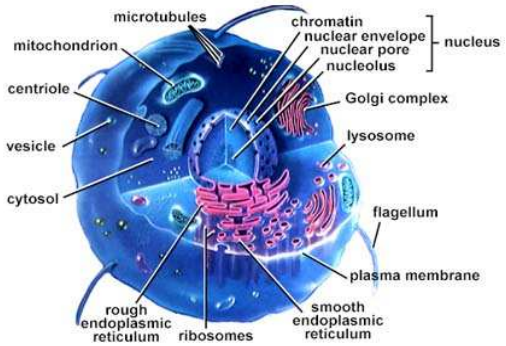
Molekula je částice složená z atomů nebo iontů. Jednotlivé části molekuly drží pohromadě síly, nazvané chemické vazby. Molekuly jsou základní stavební kameny, z nichž jsou vybudována hmotná tělesa.



Obrázek 1: *Krátký úsek molekuly DNA – dvojité spirály. Barevně jsou označeny různé druhy atomů, z kterých se skládá (vodík, uhlík, dusík a kyslík). [1]*

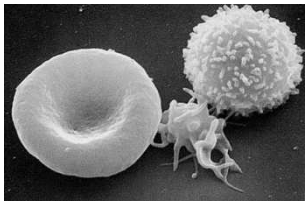
Buněčná úroveň

Buňka je složená z molekul a tvoří základní funkční jednotkou živých organismů. Zatímco některé organismy jsou pouze jednobuněčné (např. **bakterie**), jiné organismy jako člověk jsou mnohobuněčné a jejich těla se skládají z velkého počtu specializovaných buněk. Každý organismus je z buněk složen, nebo je na jiných buňkách existenčně závislý. Příkladem existenční závislosti organizmů na buňce jsou **viry**. Žádná buňka nemůže vzniknout jinak než z buňky a mateřská buňka předává dceřiné buňce potřebnou děděnou informaci k reprodukci sebe sama i ke své funkci. Buňky jsou v podstatě složeny ze stejných funkčních komponent (mitochondrie, jádro, sekreční granula, atd.).



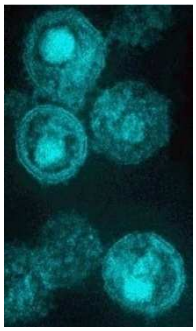
Obrázek 2: *Model živé buňky. [2]*

Konkrétní typy se liší modifikacemi pro svůj účel. Příkladem specializovaných buněk může být neuron, nebo krevní buňky.



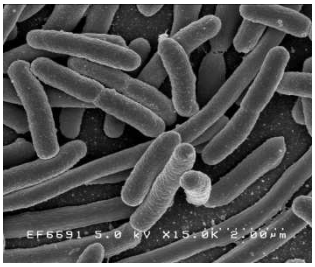
Obrázek 3: *Příklad specializovaného typu buněk, konkrétně krevní buňky, vlevo červená a vpravo bílá krvinka. [2]*

Biologický virus je **molekula**. Přejde do **buňky** a vlastním mechanizmem ji přiměje, aby místo kopírování svých vlastních molekul DNA kopírovala molekulu viru. Když se rozmnoží, protrhnou viry buňku a nakazí další buňky.



Obrázek 4: *Virus HIV. Jeho štít (povrchová membrána na fotografii) jej činí pro imunitní buňky téměř neviditelným. [4]*

Bakterie jsou jednoduché organismy. Jejich buňky jsou charakteristické přítomností buněčné stěny, nukleoidem (jadernou oblastí), specifickým typem ribozomů, DNA a také mimo jiné absencí klasického pohlavního rozmnožování.



Obrázek 5: *Bakterie Escherichia coli pod elektronovým mikroskopem. [3]*

2.1.1 DNA

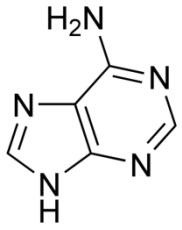
Deoxyribonucleic acid, ve zkratce *DNA* (česky deoxyribonukleová kyselina, zkratka DNK) je nositelkou genetické informace všech organismů s výjimkou těch nebuněčných, u nichž hraje tuto úlohu **RNA** (např. **viry**). DNA je nezbytnou látkou, která ve své struktuře kóduje buňkám jejich program. Tím předurčuje vývoj a vlastnosti celého organismu. U organizmů jako rostliny a živočichové je DNA uložena vždy uvnitř buněčného jádra. Jsou však i případy (např. **bakterií)**, kdy je možné, že se DNA nachází volně v cytoplazmě, tekutém prostředí buňky.

DNA je biologická **molekula**. Dvoušroubovice tvořená dvěma řetězci *nukleotidů* (stavebních kamenů kyselin v jádře) v obou vláknech. Jednotlivé nukleotidy se skládají ze tří složek:

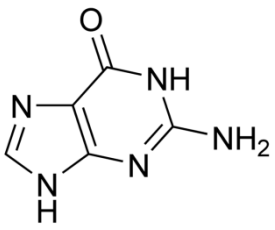
- **fosfátu** (vazebný zbytek kyseliny fosforečné)
- **deoxyribózy** (pětiuhlíkový cukr neboli pentóza)
- **nukleové báze** (konkrétní dusíkaté sloučeniny).

V DNA se v různých kombinacích vyskytují čtyři nukleové báze rozdělené do dvou skupin. U prvních dvou je základem *purin*. V čistém stavu je to krystalická zásaditá látka. Její deriváty jsou součásti nukleových kyselin. Druhé dvě jsou deriváty *pyrimidinu*. Nukleové báze jsou také jeho deriváty.

- **purinové** báze jsou:

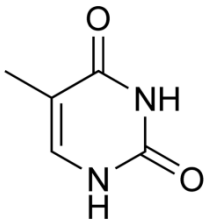


adenin (**A**)

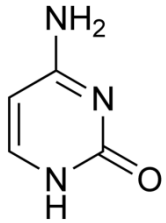


guanin (**G**)

- **pyrimidinové** báze jsou:

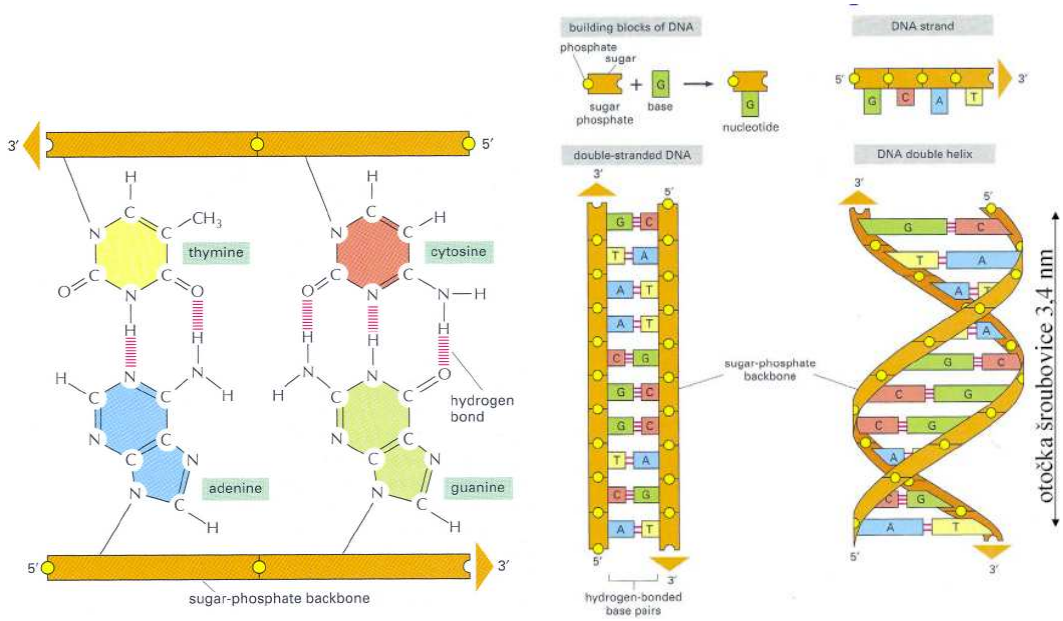


thymin (**T**)



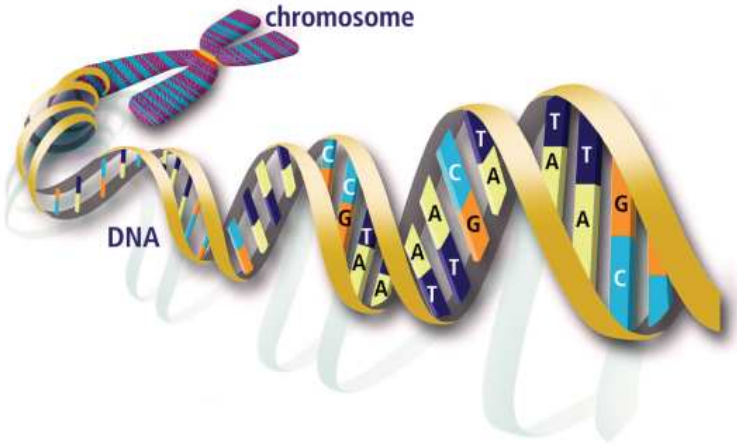
cytosin (**C**)

Tyto čtyři sloučeniny se dále na sebe váží a utvářejí páry (dvojice) v rámci svých skupin. Vazby látek mezi skupinami jsou neslučitelné. Páry se následně formují do vyšších struktur - šroubovice. Po šroubovici následuje konfigurace do chromozómu.



Obrázek 6: *Párování bází DNA [5]*

Chromozómy jsou specificky barvitelné struktury dobře pozorovatelné světelným mikroskopem. Sestávají se z DNA, RNA, histonů a jejich existence má usnadnit rovnoměrné rozdělení genetické informace do dceřiných buněk. Soubor všech chromozómů v jádře se nazývá *karyotyp*. *Histony* se podílejí na uspořádání DNA v chromozomu do vlákna vyššího řádu.

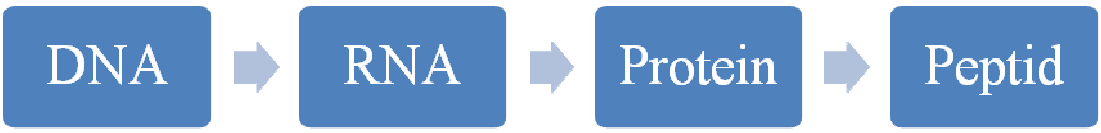


Obrázek 7: *Struktura pro složení chromozomu. [6]*

Ribonucleic acid, ve zkratce *RNA* (česky ribonukleová kyselina, zkratka RNK) je kyselina, chemicky rozlišitelná od DNA. Jednou z nejpodstatnějších funkcí

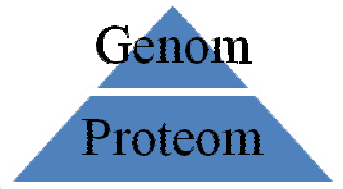
RNA je okopírovat genetickou informaci z DNA (transkripce) a fyzicky ji přenést do místa, kde po té dojde k jejímu přeložení (translace) na výsledný **protein**. *Proteiny* (česky bílkoviny) jsou molekulární přírodní látky, složené z **aminokyselin**. *Aminokyselina* je obecně jakákoliv molekula obsahující karboxylovou (-COOH) a aminovou (-NH2) skupinu. V molekulární biologii se aminokyselinou rozumí 20 konkrétních kyselin, základních stavebních složek všech proteinů. Jako příklad Alifatické skupiny se uvede alanin, leucin, valin. Z aromatické skupiny je to například fenylalanin. V polární skupině glycin, glutamin. Bazické aminokyseliny jsou lysin a arginin. Z hotových proteinů je možné vyjmenovat například albumin, hemoglobin, kolagen a lepek.

Peptid je chemická sloučenina organického původu, která vzniká spojením několika aminokyselin (obě kyseliny navázány na stejném atomu uhlíku), takzvanou peptidovou vazbou. Látky tvořené více aminokyselinami, u kterých se projevuje polymerní charakter se nazývají polypeptidy. Polypeptidy o vysoké molekulové hmotnosti se nazývají bílkoviny.



Obrázek 8: *Hierarchie sekvencí z hlediska nesené informace.*

2.1.2 Genomika a Proteomika



Obrázek 9: *Hierarchie genomu a proteomu.*

Genom je veškerá genetická informace uložená v DNA (u některých virů v RNA) konkrétního organismu. Přečtené genomy zahrnují databázi genů od „průměrného jedince“ daného druhu, protože k analýze jsou použity vždy vzorky z mnoha jedinců. Genom je tedy jednodušeji řečeno znalost úplné genetické informace organismu. *Genomika* je pak vědní disciplína s cílem zmapovat a rozkódovat genom. Genová výbava **u člověka** obsahuje přibližně $3{,}2.10^9$ [bp] vazebných párů v molekule DNA. Ty vytvářejí svými kombinacemi kolem 30 000 genů. Každá buňka aktivuje v každé chvíli určitou podmnožinu této sady. Výsledkem je obrovské množství možných stavů buněk, i když geny můžou být pouze aktivovány nebo deaktivovány. Samotné geny u jednotlivých organizmů jsou vybrané sady z mnoha možných sekvencí DNA. [7] Lidský genom byl již rozluštěn, nyní se pracuje na dekódování proteomu. Genetická informace vypovídá o převzatých schopnostech organismu. K realizaci životních funkcí je nutné vytvářet proteiny. *Proteom* je souhrn *všech* proteinů, které jsou v buňce (nebo viru), spolu s jejich interakcemi a

funkčními vztahy (tzv. specifické souvislosti). Pochází ze slov *PROTeins expressed by a genOME*. *Proteomika* je vědní obor, zabývající se globálním hodnocením genetické informace na úrovni bílkovin (proteomu). Také zkoumá strukturu a interakce proteinů.

Genom je oproti proteomu statický. Koncentrace a složení proteinů v těle se neustále mění z hlediska aktuálního stavu organizmu.

2.2 Bioinformatika

Bioinformatika je konceptualizovaná molekulární biologie. Je v ní masivně aplikována informatika (resp. informační technologie) a statistika. Cílem bioinformatiky je popsat a porozumět mj. genetickým informacím souvisejících se zkoumanými molekulami. *Genetická informace* je jednotka genetických dat ve formě genetické sekvence. *Genetická sekvence* je posloupnost písmen představujících primární strukturu reálné nebo hypotetické molekuly či vlákna DNA, které má kapacitu nést informaci. Ve formě posloupnosti písmen určující jednotlivé aminokyseliny jsou i proteomická data.

2.3 Sekvence aminokyselin a DNA

Sekvenční data jsou definována jako jakákoliv forma zápisu lineární posloupnosti *monomerů* (nízkomolekulárních látek) v molekule biologické *makromolekuly* (rozsáhlá molekula s velkou molární hmotností), nejčastěji sekvencemi DNA, nebo proteinu. Při strojovém zpracování, se na sekvenci DNA nahlíží jako zápis posloupnosti jednoznačných znaků odpovídajících jednotlivým typů monomerů po řadě tak, jak jsou řazeny v molekule při postupu ve směru odpovídajícím gradientu biosyntézy daného typu molekuly. Standardně se k zápisu používají jednopísmenové kódy monomerů tak, jak byly stanoveny Mezinárodní unií pro čistou a aplikovanou chemii (IUPAC). Pro zápis aminokyselin se také vzácně vyskytují i třípísmenové zkratky. Příklad zápisu nukleotidů a aminokyselin, viz Tabulka 2:2 a Tabulka 3:3. Digitalizací výsledků experimentálního stanovení sekvence je možné vyjádřit jedinou sekvencí *konsensus* (tj. způsob rozhodování a chování) libovolné populace molekul. S tím kódy IUPAC počítají a obsahují také možnost zápisu nejednoznačných pozic, viz Tabulka 1:1. Zápis sekvence DNA tak může obsahovat skoro celou abecedu.[8] V tomto zdroji jsou uvedeny tabulky i s více položkami.

Tabulka 1: *Označení mezer v sekvenci.*

Speciální znak	Význam
.	mezera
_	
-	

Tabulka 2: *Výběr zápisu zkratek nukleotidů v kódu UPAC.*

Reziduum DNA, RNA		
Kód	**báze**	**Pozn.**
A	Adenin	
C	Cytosin	
G	Guanin	
T	Thymin	
U	Uracil	
R	A, G	Purine
Y	C, T	Pyrimidine
B	C, G, T	Not A
D	A, G, T	Not C
...	...	...
N	Cokoliv	Any

Tabulka 3: *IUPAC výběr kódů pro zápis sekvencí aminokyselin*

Protein		
Kód	**Zkratka**	**Aminokyselina**
A	Ala	Alanin
C	Cys	Cystein
D	Asp	Asparát
...	...	...
X	Xxx	Cokoli

Kromě zmíněného zápisu sekvencí se zde vyskytují i další typy dat. Jako příklad lze uvést:

- mapy vyjadřující polohu úseků v rámci sekvence,
- vzájemná přiřazení sekvencí (lignments) a vzorce (patterns),
- údaje o 3D struktuře proteinů,
- informace o výsledcích z transkriptomických experimentů (microarrays).

Tato diplomová práce se však soustředí na data typu sekvence.

Formáty datových souborů pro zápis sekvencí

Uvnitř živých buněk jsou sekvence zapsány v posloupnosti monomerů spojených do dlouhých lineárních molekul. Aby se s těmito sekvencemi dalo pracovat, musí se sekvenční informace přenést do posloupnosti znaků, kterou lze zpracovat strojově. Realizace je ve většině případů prostřednictvím *sekvenováním* (analýza do formy řetězce ACGT jednotek) DNA. Tento proces je zpracováván v přístrojích nazývaných sekvenátory. Sekvenční informace je obecně jako každá informace nehmotná a molekula DNA se plní funkci média či nosiče, na němž je

sekvence zapsána. [9] DNA je hmotnou molekulou, která reaguje s řadou látek a v průběhu *sekvenování* se a ne vždy chová ideálně. Běžné metody sekvenování jsou založeny na vytváření souborů definovaných krátkých fragmentů DNA a jejich *elektroforetickém* dělení. Elektroforéza je soubor separačních metod, které využívají k vytváření látek jejich odlišnou pohyblivost ve stejnosměrném elektrickém poli. Dělení probíhá s velkou přesností. Elektroforéza pracuje s velkými populacemi molekul, které se chovají statisticky. Proto generuje data analogová, nikoli digitální. Chování fragmentů DNA v průběhu dělení odráží navíc kromě jejich velikosti i tvar. To se projevuje ve snaze vytvářet stabilní sekundární struktury. [8]

V nejjednodušší digitalizované podobě je sekvence zapsána jako prostý řetězec IUPAC znaků v textovém souboru. Tento formát se označuje jako *surová data* (raw data, raw formát). Výhodou je udržet sekvenci pod limitem 255 znaků. Formát surových dat se pro dlouhodobé uchování (archivaci, zálohování) nehodí, protože neumožňuje uložit dodatečné informace (hlavičky) jinam, než do názvu souboru. Pro uchovávání sekvenčních dat spolu s dalšími průvodními informacemi byla vyvinuta řada formátů. Specializované databáze jako EMBL, GenBank a DDBJ používají několik vzájemně převoditelných formátů, které poskytují prostor pro poznámky. [8]

Formát FASTA

Jedním z nejpoužívanějších formátů datových souborů má název **FASTA**. Jedná se o textový soubor, jehož první řádka začíná znakem > (větší než). Za tímto znakem následuje *hlavička* obsahující název sekvence, anotaci a případně další údaje, které nejsou součástí samotné sekvence. Hlavička se dělí na skupiny znaků. První skupina alfanumerických znaků (až po první mezeru) představuje jedinečný identifikátor sekvence (název genu či klonu). Další skupina je volitelný komentář. Za hlavičkou následuje surová sekvence. Programy zpracovávající sekvence ve formátu FASTA obsah hlavičky ignorují, popřípadě používají několik prvních znaků (název sekvence). Samotná sekvence by neměla obsahovat prázdné řádky a mezery. [8]

Dalším příkladem jednoduchého zápisu je formát **PIR/NBRF** (Protein Information Ressource/National Biomedical Research Foundation). Je podobný formátu FASTA, nikoliv však kompatibilní. Je zde uveden pro upozornění na možnou záměnu. [8]

Existuje řada volně dostupných softwarových nástrojů pro formátování sekvencí a vzájemné převody mezi jednotlivými formáty. Soubor nástrojů pro převod sekvencí nejrůznějších formátů na FASTA je například součástí freewarového balíku *The Sequence Manipulation Suite* – SMS. [8]

2.4 Zdroje genetických dat

Volné sdílení sekvenčních dat je běžné v elektronických databázích. Významnější časopisy požadují, aby byla data, o kterých je článek, byla volně přístupná veřejnosti. V případě sekvencí DNA se toto realizuje prostřednictvím mnoha přístupných veřejných databází. Přehled aktuálního stavu je každoročně zveřejňován ve specializovaném čísle časopisu Nucleic Acids Research. [8]

Rozdělení

Obecně se rozlišují databáze *moderované* (cuated) a *nemoderované*. Do nemoderovaných databází může přispívat kdokoliv. Zveřejnění výsledků v této databázi je obdoba nerecenzované publikace. Na druhou stranu moderované databáze přijímají data podle kritérií stanovených správcem databáze. Příkladem této databáze je Swissprot. [8]

Databáze sekvencí

Nejvýznamnější jsou primární databáze mezinárodního konsorcia International *Nucleotide Sequence Database Collaboration.* Ta zahrnuje americkou databázi ***GenBank***, evropskou ***EMBL*** (European Molecular Biology Laboratory Data Libary) a japonskou ***DDBJ*** (DNA Data Bank of Japan). Tyto tři databáze si informace denně vyměňují a tím i navzájem zálohují. Jejich obsah je v podstatě totožný. Primární databáze jsou přístupné přes webové rozhraní. Přes FTP se dá zdarma stáhnout celá databáze. Nevýhoda vyjmenovaných databází je, že jsou nemoderované. [8]

Vyhledávání v databázích

K vyhledávání a stahování záznamů z databází slouží webové uživatelské rozhranní. Na rozdíl od obsahu je toto uživatelské prostředí různé. V Evropě je nejběžnější přístup prostřednictvím rozhranní ***SRS*** (Sequence Retrieval System) z databáze EMBL na Evropském ústavu pro bioinformatiku (European Bioinformatics Institute, EBIF). Druhá možnost je rozhranní ***Entrez*** databáze GenBank v americkém Národním centru pro biotechnologické informace (National Center for Biotechnology Information). [8]

Databáze proteinové sekvence

Mimo uvedené databáze je užitečná databáze proteinových sekvencí ***Swissprot***. Dále americká databáze ***PIR*** (Protein Information Resource) a evropská ***trEMBL***. Tyto tři databáze jsou spojené pod hlavičkou iniciativy *UNIPROT* a sdílejí svá data jako skupina první (GenBank, EMBL, DDBJ). [8]

Další významná databáze nese název **IPI**. Název je zkratka slov ***International Protein Index***. Konkrétně byla použita verze 3.39 ve formátu zápisu fasta <IPI.HUMAN.v3.39.fasta>. Zdroj [10]

IPI je sestavený Evropským Bioinformatickým Institutem. Nese zkratku EBI z anglického *European Bioinformatics Institute*. EBI poskytuje hlavní databázi, která popisuje lidský (human) proteom a proteom myši. Báze tedy nese data peptidů a proteinů z člověka a myši. Další možné vzorové databáze jsou již zmíněné: SWISS-PROT, TrEMBL, NCBI RefSeq a Ensembl. Použití IPI je výhodné ve vždy aktualizovaných verzích zaručené křížovými odkazy mezi primárními zdroji dat. V IPI se také vyskytuje minimální nadbytečnost dat, což pomáhá integritě vyhledávání. IPI je aktualizována každý měsíc. [10]

2.5 Identifikace proteomu neznámého vzorku

Tato úloha jde v bioinformatice obecně dvěma cestami. Pro organismy s ***nesekvenovaným genomem*** se používá hledání na základě evoluční homologie proteinových sekvencí (přístup jako BLAST, N-W algoritmus). Tato problematika je vice popsaná ve zdroji [11]. Pro organismy se ***sekvenovaným genomem*** se uplatňuje statistická shoda. Zde se využívá konkrétně rozpoznání peptidů (např. z MS/MS hmotnostní spektrometrie) srovnávané s experimentálně získanými hodnotami v databázi (přístupy např. Mascot, X!Tandem).

2.6 Rozpoznávání proteinu s nesekvenovaným genomem

Podobnosti dvou sekvencí

Hlavním tématem bioinformatiky u analýz nesekvenovaného genomu je zkoumání podobnosti dvou sekvencí. Biologické sekvence vývojem v čase a evolucí ukázaly, že ne všechny změny v biologických sekvencích jsou možné se stejnou pravděpodobností. Určité aminokyseliny se v procesu přeměny jedné aminokyseliny na druhou přeměňují často, zatímco jiné substituce jsou velmi cenné. Založením na informaci evolučního vývoje proteinů byl odvozen model ve formě matice, nazývaná jako *matice substitucí* (substitution matrix, scoring matrix - matice ohodnocení).

Pro sekvence nukleových kyselin se používají varianty *jediné matice*. Ta nese název **IUPAC,** matice identity (identity matrix)**.** Ta všem souhlasným párům přiřazuje konstantní hodnotu kladnou a nesouhlasným párům hodnotu zápornou, nebo rovnu nule. [8]

Matice podobnosti PAM

U proteinů je také možné vybrat z několika druhů matic. Jedna z nich má název **matice PAM** (Point Accepted Mutation). První matice PAM byla publikovaná v roce 1978 Dayhoffem. Je sestavena z globálně zarovnaných sekvencí s podobností 85%. PAM matice udává pravděpodobnost toho, že jakákoliv daná aminokyselina zmutuje do jiné v daném časovém intervalu. Například PAM1 říká, že 1 aminokyselina ze 100 bude zmutována v daném časovém úseku. Na druhé straně měřítka matice PAM256 udává pravděpodobnost 256 mutací ve 100 aminokyselinách. Databáze znalostí, na které byla sestrojena první PAM matice je v dnešní době zastaralá. V algoritmu se předpokládá, že všechny aminokyseliny mutují se stejnou rychlostí, což ovšem není správný předpoklad.

Konstrukce vychází z empirického stanovení frekvence jednotlivých specifických záměn. V rámci modelové sady sekvencí byly pro každou aminokyselinu zjištěny frekvence všech možných záměnových mutací v situaci, kdy tyto mutace postihují maximálně 1% zbytku v sekvenci. Na základě takto změřených hodnot specifických mutačních rychlostí byla sestrojena matice PAM1. Vynásobením této matice jí samotnou vzniká matice PAM2. Takto se pokračuje do matice PAM250, která odpovídá stavu, kdy na 100 aminokyselinový úsek připadá 250 mutací. Mutace postihuje náhodně vybrané zbytky. Pořád zde zůstává zachováno 20% zbytků v pořádku (nemutovaných). Tato hodnota je brána za limitní mez. [8]

Obdobou série PAM jsou matice **JTT** (Jones-Taylo-Thorton) a **Gonnetovy** matice. Pro číslování těchto matic platí totéž co pro PAM, jen jsou sestrojeny z většího souboru výchozích dat. GONNET250 je ekvivalentní PAM250. PAM matice pro vzdálenější frekvence (s vyššími hodnotami v matici) jsou získány extrapolací. [8]

Tabulka 4: *Matice PAM120*

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

Matice podobnosti BLOSUM

V roce 1992, 14 let po publikaci první PAM matice, vzniká matice BLOSUM (BLOcks SUbstitution Matrix). Ta byla vyvinuta a publikována v knize Stevena Henikoffa. [9][40] Henikoff a kolektiv hledal jako základ pro svůj model více odlišné proteiny. Použil lokálně zarovnané sekvence, kde žádná z použité sekvence nebyla s totožností menší než 60%. Tento výsledek dal vzniknout matici nazvanou BLOSUM62. Ve srovnání s PAM maticemi jsou matice BLOSUM vypočteny ze zarovnání bez mezer objevující se v databázi bloků. Sean Eddy v roce 2004 popsal substituční matici BLOSUM62 a postup jak ji sestavit. [39] Matice **BLOSUM** se vyskytují v sériích. Ta je odvozena z empirických dat na základě lokálního přiřazení konzervativních domén méně blízce příbuzných sekvencí. Také matice BLOSUM jsou číslovány obdobným systémem. Čím vyšší číslo, tím vyšší frakce zbytků ve výchozím souboru sekvencí. Vysoké BLOSUM se volí tam, kde je předpoklad podobnosti v lokálních blocích (příčinou selekce, nebo evolučním omezením). [8]

Matice BLOSUM se standardně používají pro široké spektrum zarovnávacích programů. Pro příklad je zde uvedena jedna ze série, matice s číslem 62 (označení Blosum62), viz Tabulka 5:. V tabulce je zeleně vyznačena polovina matice, která je symetrická podle hlavní diagonály (žlutě vyznačené prvky). Šedá pole představují pomocné výpočty. Matice byla vygenerována funkcí příkazem *blosum(62)* v prostředí MATLAB verze 7.2.0.232 (R2006a) využívající *Bioinfomatics toolbox*. Hodnocení (skóre) pro hledaný uspořádaný pár odpovídá kombinaci souřadnicím řádků a sloupců. Například substituční skóre (shodnost) pro arigin (R) a lysin (K) je 2. V matici, zobrazené tabulce je označeno červenou barvou. Diagonální prvky matice představují aminokyseliny beze změn, tzn. kdy zarovnaný pár je tvořen dvěma totožnými aminokyselinami. Většina substitucí je ohodnocena má zápornou hodnotou. Platí, že čím menší ohodnocení, tím je menší pravděpodobnost shody. V matici se vyskytují pouze zaokrouhlené hodnoty.

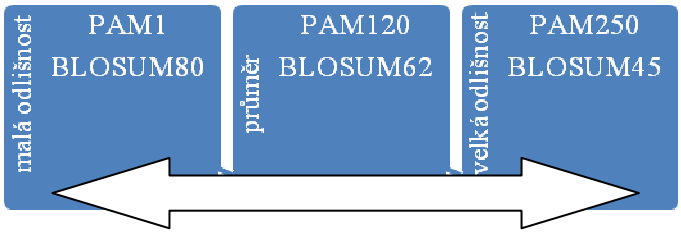
Tabulka 5: *Matice Blosum62 - přehled kompletní matice zahrnující všechny možné substituční kombinace a jejich ohodnocení. Hlavička „*“ je hodnota mezery.*

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Srovnání PAM a BLOSUM

Čím podobnější jsou si sekvence proteinů, tím se volí nižší hodnoty PAM a vyšší hodnoty BLOSUM. Matice s nízkým PAM a vysokým BLOSUM nejsou plnohodnotně zaměnitelné, protože vznikly z jiných empirických dat. Nízké PAM se volí kdy je mezi sekvencemi krátká evoluční vzdálenost (malý počet mutací). [40] [40]

Tabulka 6: *Vztah mezi skórovacími maticemi BLOSUM a PAM.*



Typy zarovnání

Je možné obecný postup porovnání dvou sekvencí rozšířit na sekvence velmi odlišné, nebo různé délky (důsledek delece, či inzerce). Jedna z možností je přiřazení sekvencí i za cenu zanesení mezer do jedné, či obou z páru porovnávaných sekvencí. Tento přístup se nazve *globální přiřazení*. Efektivní pro sekvence, které jsou si navzájem *blízké*. Blízkou se rozumí sekvence, která nepodlehla v průběhu evoluce přestavbám a lze je vzájemně přiřadit vnesením malého počtu mezer. Nebo zda se omezí porovnávání na úseky sekvencí tam, kde jsou vzájemně podobné na únosnou míru a nepřiřazené úseky se budou ignorovat – *lokální přiřazení*. [8]

Tabulka 7: *Příklad globálního (nahoře) a lokálního (dole) přiřazení při zarovnávání dvou sekvencí.*

S1	M	A	R	T	I	N	B	L	A	H	O	N	Z	A	B	L	A	B	L	U	C	K	A
S2	M	A	R	T	I	-	-	-	-	-	-	N	-	A	-	L	-	-	-	U	C	K	A

S1	S	N	A	V	N	A	P	A	T	N	I	K	U	A	P	I	O	N	Y	R	R	A	K
S2	-	-	-	-	N	A	P	A	T	N	I	K	U	-	P	I	O	N	Y	R	-	-	-

2.7 Rozpoznávání proteinu se sekvenovaným genomem

Rozpoznání je rozděleno do dvou úrovní. První je fyzické zpracování neznámého vzorku (např. metodou MS) a druhá identifikace výsledků s matematickým aparátem. Jedna z metod použitelná ke studiu proteomu je *tandemová hmotnostní spektrometrie*.

2.8 Tandemová hmotnostní spektrometrie

Zkoumané vzorky určené k rozpoznání, je potřeba nejprve analyzovat na přístroji s názvem *hmotnostní spektrometr*. Na tomto přístroji se realizuje metoda

s názvem *hmotnostní spektrometrie*. Po rozšíření metody do spolupráce více analyzátorů, tzv. tandemu, vznikne metoda zvaná *tandemová hmotnostní spektrometrie* realizovaná na *tandemovém hmotnostním spektrometru*.

Každý vzorek látky obsahuje škálu četnosti komponent, ze kterých se vzorek skládá. Přístroj provádí analytickou techniku orientovanou na určení iontů v látce, přesněji separuje ionty podle poměru jejich hmotnosti a náboje (m/z). Získaný záznam je spektrum hmotností analyzované látky (analytu). Tento postup je nejvíce používán k analýze složení fyzického vzorku. A tím pro identifikaci neznámé směsi sledované látky.

2.8.1 Hmotnostní spektrometr

Metoda hmotnostní spektrometrie se realizuje na přístroji zvaný hmotnostní spektrometr. V anglickém originále *mass spectrometer (zkratka - MS).*

Obrázek 10: *Hmotnostní spektrometr firmy Thermo, model DECA XP. [41]*

Komponenty přístroje

Ionizováním vzorku se dosahuje oddělení jednotlivých iontů. Obecně je přístroj složen ze tří částí: zdroj iontů, kumulační analyzátor a detektor.

Podrobněji se dá hmotnostní spektrometr popsat takto:

- Zdroj iontů – převedení do plynného skupenství a ionizace plynu hledané látky.
- Hmotnostní analyzátor – oddělení iontů vzhledem k hmotnosti, řídí se Newtonovými zákony a Lorenzovou silou.
  - Lorenzova síla: $F = q(E + v \times B)$, a Newtonova síla: $F = m \cdot a$, kde
    - F – síla aplikovaná na iont
    - m – hmotnost iontu
    - q – iontový náboj
    - E – elektrické pole
    - $(v \times B)$ je vektorový produkt rychlosti iontu „v“ a magnetického pole „B“

- Sektor – ovlivňuje cestu iontů elektrickým, nebo magnetickým polem
- Doba života v letu iontu (Time of Flight, zkráceně TOF) – analyzátor zrychlující ionty přes stejný elektrický potenciál a měřící jejich časy při letu. Rychlosti závisí na jejich hmotnosti. Lehčí ionty jsou v cíli dříve.
- Pasti (anglicky traps), alternativa k TOF
    - Kvadratické pole (QMS) – používají kmitavá elektrická pole, která selektivně stabilizují, nebo naopak destabilizují ionty svojí radiovou frekvencí (RF puls, princip používán v nukleární magnetické rezonanci).
    - Kvadratická Iontová past (QIT) – pracuje stejně jako QMS, jen že ionty jsou zde zachycovány a následně katapultovány.
    - Lineární kvadratická iontová past (LQIT) – Obdobná QIT, jen místo 3D pasti je zde past 2D.
    - Fourier
    - Orbitrap
- Detektor – poslední součást hmotnostního spektrometru. Zaznamenává náboje iontů v poměru jejich hmotností (m/z).

Funkce přístroje

Různé chemikálie mají různou hmotnost. Tato skutečnost je využita v analýze vzorku. Jako příklad se uvede chlorid sodný (NaCl), tedy kuchyňská sůl. V prvním kroku je odpařena a tím se změní na plyn. Plyn je ionizován (z funkčního hlediska fragmentován) do elektrický nabitých částic ze kterých byla látka složena (iontů). Ionty sodíku a chlóru mají specifickou hmotnost atomu (atomovou hmotnost). Také je zde možnost zachytávat buď elektrickým, nebo magnetickým polem.

Všechny ionty jsou odeslány do iontové akcelerační komory. Zde se aplikuje magnetické pole. To odchýlí iont každého prvku jinou cestou. Lehčí ionty se odchylují dále než těžké, protože síla každého iontu je úměrná jeho hmotnosti. Proces se chová dle rovnice $F = m \cdot a$. Z vzorce vyplývá, že pokud síla a hmotnost zůstane konstantní, zrychlení se bude měnit s nepřímou úměrou.

Ionty jsou odchýleny přímo na detektor. Detektor změří, jak moc byl iont odchýlen a z tohoto vychýlení vypočítá s vysokou pravděpodobností jeho hmotnost a po skončení měření všech iontů chemické složení analyzované látky.

2.8.2 Způsoby ionizace

Vstupem je látka v jakémkoli skupenství. Iontový zdroj hmotnostního spektrometru slouží k převedení neutrálních molekul analytu na nabité částice (ionty). Je potřeba více ionizačních technik, protože se látky mohou lišit podle mnoha různých chemických hledisek. Je třeba vybrat optimální způsob ionizace pro danou látku. Pro analýzu biomolekul je vhodná ionizační technika ESI a MALDI. [12]

ESI

Zkratka plynné ionizace, uskutečněná většinou ionizujícím sprejem (elektrosprejem) aplikovaným na analyzovanou látku. Analyzují se většinou ionizované plyny měřené látky.

MALDI

Dosavadní způsoby ionizace a detekce umožňovaly analyzovat látky jen s nízkou molekulovou hmotností. Kromě plynné, existují i další ionizační techniky.

K stanovení vyšších molekulových hmotností se používá ionizace laserem za přítomnosti matrice (MALDI, matrix assisted laser desorption/ionization) v kombinaci s detektorem doby letu (TOF, time-of-flight). Detektor umožňuje změřit dobu průletu a z ní lze vypočítat rychlost částice. Ionty analyzované látky jsou urychleny silným elektrickým polem (25–30 kV) a přes uzemněnou mřížku vstupují do vakua v trubici detektoru letu, kde se pohybují rychlostí danou jejich hmotností a nábojem. Zde se měří doba letu částice, z níž se pak vypočte poměr molekulové hmotnosti a náboje částice.

Hmotnostní spektrometrie „MALDI“ je metoda, která byla původně vyvinuta pro kvalitativní analýzu peptidů a bílkovin, avšak nyní se využívá i pro analýzy nukleových kyselin nebo nízkomolekulárních organických i anorganických látek. Výhodou je vysoká citlivost a rychlost měření. Metodu MALDI lze kombinovat jako hmotnostní detektor se separačními metodami, například gelovou elektroforézou či kapalinovou chromatografií. K hlavním cílům patří možnost využití při kvantifikaci analyzované látky.[13]

2.8.3 Základní typy hmotnostních analyzátorů

Magnetický analyzátor – v magnetickém poli dochází k zakřivení dráhy iontu závislé na hodnotě m/z. **Analyzátor doby letu (TOF)** – urychlené ionty se v oblasti bez pole pohybují různou rychlostí v závislosti na hodnotě m/z (“menší ionty letí rychleji“). **Kvadrupólový analyzátor –** v daný okamžik jsou oscilace stabilní pouze pro určitou hodnotu m/z a tento ion projde kvadrupólovým analyzátorem, ionty s jinými hodnotami m/z mají nestabilní oscilace a jsou zachyceny na tyčích kvadrupólu, změnou napětí postupně projdou na detektor ionty se všemi hodnotami m/z. **Iontová past –** trojrozměrná analogie kvadrupólu, opakované selektivní vypuzení iontů podle hodnoty m/z z iontové pasti na detektor, možnost MS analýzy. [12]

2.8.4 Tandemový hmotnostní spektrometr

Tandemový hmotnostní spektrometr, zkráceně MS/MS nebo TMS, je jeden z typů analytických přístrojů známých jako hmotnostní spektrometry. Hmotnostní spektrometr měří hmotnost molekul v ionizovaném stavu. V tandemovém hmotnostním spektrometru jsou umístěny dva hmotnostní spektrometry v páru a mezi nimi dochází k rozbití analyzované molekuly na fragmenty. Podle jejich spektra lze následně látky identifikovat. Strukturně podobné látky se v některých fragmentech shodují, např. acylkarnitiny mají společný fragment o hmotnosti 85 Da, aminokyseliny fragment o hmotnosti 102 Da. To je možné využít pro jejich stanovení ve složité směsi (biologický materiál) bez použití separačních technik. [14]

TMS se dá tedy představit jako více MS spolupracujících analyzátorů. To umožňuje vícenásobné kroky výběru a analýzy. Například, jeden MS analyzátor může izolovat jeden peptid od ostatních. Druhý MS analyzátor pak stabilizuje ionty peptidu během srážky s plynem (kolizní fragmenty se nazývají CID, *Collision*

Induced Dissociation). Třetí hmotnostní analyzátor katalogizuje fragmenty peptidů. MS/MS tandemová spektrometrie může být také realizována v jednom hmotnostním analyzátoru.

Další příklad může být, že první TOF analyzátor (prvního TMS) vybírá ionty, které vstupují do cely umístěné mezi TOF analyzátory. Spektrum fragmentů je měřeno druhým TOF analyzátorem (druhý TMS). [15]

2.8.5 Hmotnostní Spektrometrie

Hmotnostní spektrometrie (MS) je analytická metoda sloužící k převedení molekul na ionty, rozlišení těchto iontů podle poměru hmotnosti a náboje (m/z) a následnému záznamu relativních intenzit jednotlivých iontů. Je tedy založena na rozdělení nabitých částic podle jejich molekulových hmotností. Mimořádně citlivá, destruktivní a s minimální spotřebou vzorku. [12]

Hmotnostní spektrometrie MS je metoda umožňující měření hmotnosti látek. Díky zjištění molekulární hmotnosti prvků obsažených v látce (proteinu) umožní její identifikaci. Identifikace proteinu se provádí dvěma základními způsoby.

Prvním způsobem je použití enzymu pro štěpení proteinu na menší peptidy, jejichž přesné hmotnosti jsou pomocí MS změřeny. Spektrum těchto hmotností je pak porovnáno s teoretickými spektry, která jsou vypočítána ze sekvencí proteinů v dostupných databázích.

Druhým je Tandemová MS (TMS, nebo MS/MS). Ta umožňuje zvolit peptid, který je následně fragmentován. Profil výsledků fragmentace poskytuje informaci o sekvenci proteinu, která je srovnána s daty uloženými v databázích. [16]

2.8.6 Tandemová hmotnostní spektrometrie

Tandemová hmotnostní spektrometrie, ve zkratce TMS, nebo MS/MS. Z anglického *tandem mass spectrometry*. Tandemová MS zahrnuje vícenásobné kroky hmotnostní analýzy, obvykle separování formou fragmentace. Tandemový hmotnostní spektrometr je schopný vícenásobných dějů. Například jeden analyzátor hmotností může izolovat jeden peptid z mnoha, které vstupují do spektrometru. Druhý hmotnostní analyzátor potom stabilizuje ionty peptidů, dokud se nesrazí s plynem. V tomto případě se jedná o kolizní separování, anglicky *collision-induced dissociation* (CID). Po fragmentaci, třetí hmotnostní analyzátor katalogizuje fragmenty vyprodukované vstupními peptidy. Tandemová MS může být zrealizována v jednom hmotnostním analyzátoru, to je možné s čtyřpólovou iontovou pastí (quadrupole ion trap). Jsou zde i jiné techniky pro fragmentaci molekul tandemovou MS. Mimo již zmíněnou CID existuje i *electron capture dissociation* (ECD), *electron transfer dissociation* (ETD), *infrared multiphoton dissociation* (IRMPD) and *blackbody infrared radiative dissociation* (BIRD). Důležitým využitím tandemové MS je pro aplikaci identifikace proteinů.

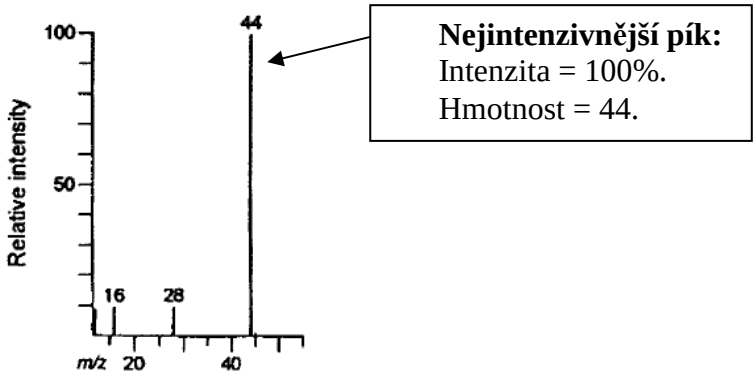
2.9 MS/MS data

Hmotnostní spektrum

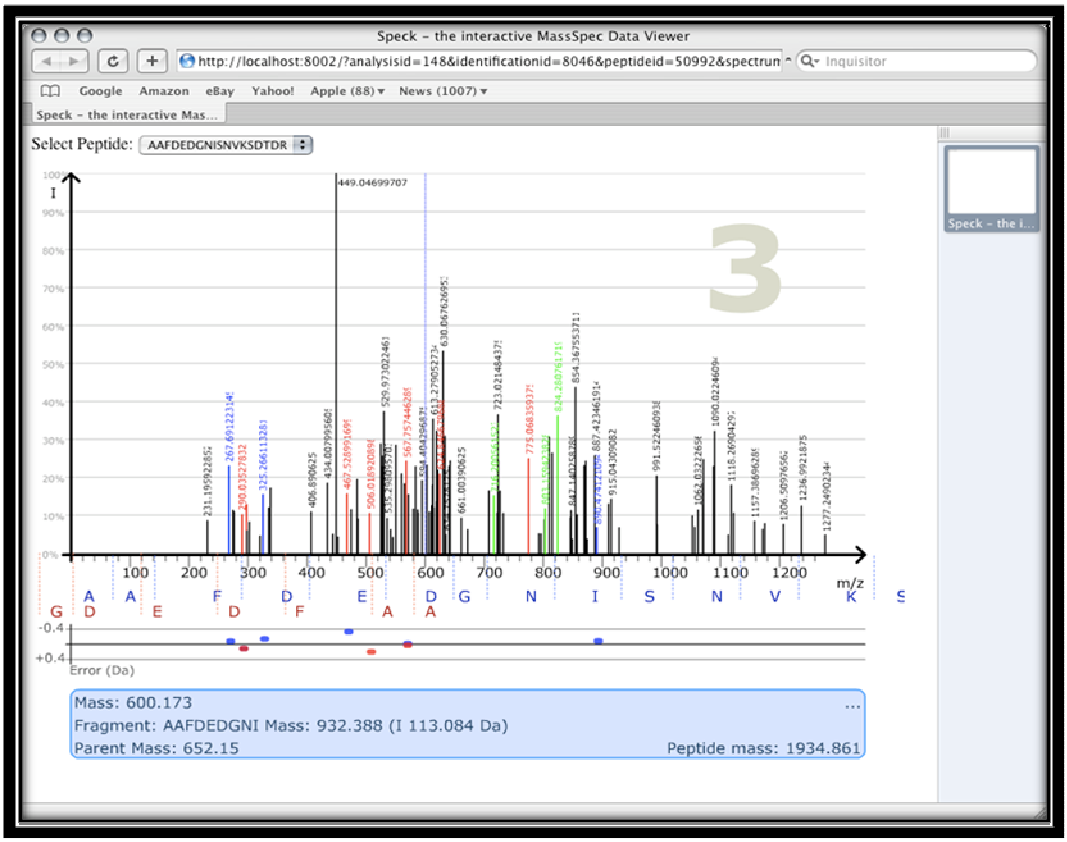
Hmotnostní spektra jsou obvykle uváděna v *normalizovaném tvaru*, tzn. nejintenzivnějšímu píku spektra je přiřazena relativní intenzita 100% a intenzity

ostatních píku se dopočítají. Kromě grafické formy lze spektrum uvést i v tabelární formě (přesně uvedené intenzity, ale méně přehledné pro interpretaci). [12]

- osa y = relativní intenzita v %,
- osa x = poměr hmotnosti a náboje (m/z), ve většině případů je náboj roven jedné (kromě ESI) a potom hodnota na ose x odpovídá přímo hmotnosti iontu



Obrázek 11: *Vzorový příklad píků hmotnostního spektra.[12]*



Obrázek 12: *Výstup naměřených dat z komerčního programu Overwinter.[17]*

Databáze MS/MS dat

Velké množství zdrojů naměřených dat je přístupná na internetu. Databáze obsahuje širokou škálu změřených nukleových kyselin mnoha organizmů. Relevantní surová změřená raw data pro tuto práci, jsou označeny „Organism human“, tedy lidské vzorky. [18][19]

Standard formátu pro uložení naměřených dat

Data získaná z tandemového hmotnostního spektrometru jsou uložena v jednom z nejpoužívanějším formátu *mzXML*. Jeho podrobnější popis je ve zdroji [19] a [26], sekce schéma formátu. V práci byl použit k přehledu vstupních mzXML dat prohlížeč Insilicos od firmy Life Science Software.

Zavedený standard uložení naměřených „mz“ dat (mzXML) se ukazuje jako příliš těžkopádný. Pracovní skupina pro standardy hmotnostní spektrometrie se zabývá definicemi datových formátů a archivací v hmotnostní spektrometrii proteomu. Právě probíhá vývoj standardu formátu s označením mzML. Více informací se nalezne ve zdroji [20].

2.10 Analýza MS/MS dat

Po fyzické analýze nastupují metody programového rozpoznávání analytu. Data z MS/MS (TMS) reprezentují histogramy rozložení hmotnosti analyzované látky. Spektrum výsledných relativních molekulových hmotností je porovnáno se soubory hmotností vytvořených počítačem z proteinů v proteinové databázi. Pro identifikaci je vhodné zvolit nějaký z již ověřených algoritmů. Zde se práce zaměří na dva nejpoužívanější, Mascot a X!Tandem.

U obou (Mascot, X!Tandem) je **protein identifikován na základě *srovnání* experimentálně získané peptidové mapy s teoretickými peptidovými mapami proteinů v databázi**. Následně dochází k výpočtům generující identifikaci proteinu a stupeň věrohodnosti výsledku.

2.11 Mascot

Prvním použitým softwarovým nástrojem pro hodnocení a identifikaci. Používá data z hmotnostní spektrometrie pro určení proteinu, který se nachází v primární databázi sekvencí. Tyto shody v databázi dále zpracuje a udá míru příslušnosti, nazvanou skóre.

Tento software vyvíjí firma *Matrix Science* a je dostupný pouze komerčně. Vzhledem k ochraně firemního „know how“, se na oficiálních stránkách [10] produktu Mascot hovoří o skórovacích algoritmech obecně.

2.11.1 Schéma skórování

Hodnocení se v podstatě dělí na dvě důležité hodnoty. První je skóre vyjadřující stupeň příslušnosti přiřazeného proteinu a druhá hodnota vyjadřuje správnost výsledného skóre z hlediska pravděpodobnosti.

ION-SKÓRE

Mascot využívá základní principy algoritmu MOWSE. MOWSE je zkratka anglických slov MOlecular Weight SEarch. Přeloženo do češtiny jde o vyhledávání pomocí hmotností molekul. Schéma skórování MOWSE je podrobněji popsané ve zdroji [21]. Jde o hodnocení založené na distribuci četnosti peptidů z OWL (non-redundantní databáze). Pracovní tok je složen ze dvou stupňů vyhodnocení.

První stupeň vyhledávání je porovnání vypočtené peptidové mapy hmotností s každou položkou v databázi sekvencí (sadou experimentálně získaných dat). Každá vypočtená hodnota, která spadá do zadané tolerance hmotnosti v experimentálně získaných hodnotách databáze, je brána jako *shoda* (match). Jako primární filtr může být použit molekulární rozsah hmotností pro neporušený protein. Přesněji řečeno, se počítá počet shodných peptidů.

Mowse využívá **empiricky určené činitele pro přiřazení statistické významnosti** (váhy) pro každou shodu peptidu. *Matice faktorů (vah k výsledkům)* je spočtena během sestavovaní databáze a postup jejího výpočtu je následující: Nejprve se určí tzv. *četnostní faktor* $f_{i,j}$ matice **F** (z anglického slova frequency). Ten se vytvoří tak, že každý řádek (anglicky „row“) představuje interval 100 Da (Dalton[1]) v peptidovém spektru hmotností. Každý sloupec je v intervalu 10 kDa v neporušeném proteinovém spektru hmotností. V každém vstupu sekvence probíhá zpracování vhodné matice elementů $f_{i,j}$. Ty jsou inkrementovány tak, jak roste rozdělení pravděpodobnosti hmotnosti. Funguje tedy jako funkce spektra hmotností molekul v proteinu.

Elementy frekvenční matice F jsou potom normalizovány dělením prvků. U každého j-tého 10 kDa sloupce probíhá dělení s jeho nejvyšší hodnotou. Výsledky normalizace udávají *Mowse faktory* $m_{i,j}$“ matice **M**:

$$m_{i,j} = \frac{f_{i,j}}{|f_{i,j}|_{\text{max ve sloupcích } j}} \tag{1}$$

V druhém stupni hodnocení (po vyhledání všech experimentálních hodnot) pokračuje výpočet hmotností peptidové databáze M_p. Nakonec je **skóre pro každý vstup** vypočteno ze vzorce:

$$Score_{i,j} = \frac{50\,000}{M_p \cdot \prod_n (m_{i,j})} \tag{2}$$

kde M_p je molekulární hmotnost vstupu, $\prod_n (m_{i,j})$ je produkt Mowse faktorů, $50\,000$ je hodnota pro normalizaci k „průměrnému“ proteinu o 50 000 kDa.

Produkt $\prod_n (m_{i,j})$, má matematický význam jako výsledek skalárního součinu. Je vypočten z Mowse faktoru pro každou shodu mezi experimentálními daty a peptidovou mapou hmotností, vypočtenou ze vstupních dat.

[1] *Dalton* je jednotka relativní molekulové hmotnosti. Značka [Da]. Udává kolikrát je klidová hmotnost daného atomu větší než dohodnutá unifikovaná atomová hmotnostní konstanta. [38]

Mowse skóre na základě pravděpodobnosti

Mascot v sobě zahrnuje implementaci Mowse algoritmu na pravděpodobnostním základě. Mowse algoritmus je například vhodný pro modelování chování proteolytického enzymu. Základním projevem proteolýzy je štěpení peptidových vazeb bílkovin a peptidů. Více informací k této problematice je možné nalézt ve zdroji [22].

Dosazením Mowse algoritmu do pravděpodobnostního rámce se získají nové výhody:

- může být použito jednoduché pravidlo pro rozhodování, zda je výsledek významný, nebo ne,
- rozdílné typy shody (hmotnostní mapy peptidů a fragmenty iontů) můžou být kombinované do jednoduchého vyhledání,
- skóre z rozdílných vyhledávání nad odlišnými databázemi mohou být porovnávány,
- vyhledávácí parametry mohou být optimalizovány snadněji a v iteracích.

Shody využívají hodnot molekulových hmotností (jedno spektrum hmotností u MS, nebo MS/MS fragmenty iontových spekter). Ty jsou vždy řízené na pravděpodobnostní úrovni. Výsledné skóre je absolutní pravděpodobnost, že sledovaná shoda je náhodný jev. Výsledné pravděpodobnosti mohou být ale matoucí. Protože zahrnují velký rozsah amplitud, je pro „vysoké“ skóre „malá“ pravděpodobnost. To je nejednoznačné. Z tohoto důvodu se ve skutečnosti počítá skóre *M* jako:

$$M = -10 \cdot log_{10}(P), \tag{3}$$

kde *P* je absolutní pravděpodobnost správné shody.

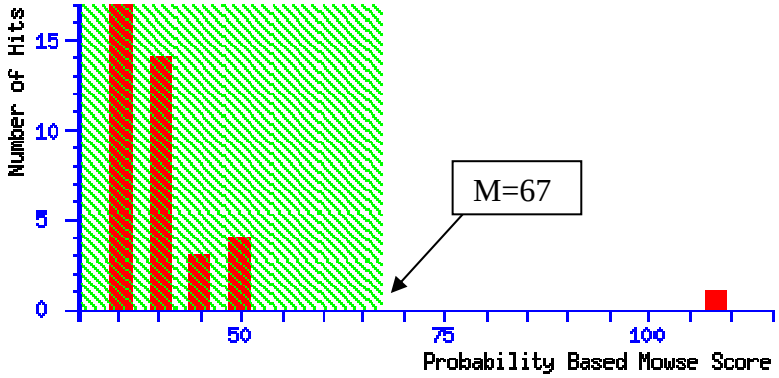
Příklad výpočtu pro pravděpodobnost: $P = 10^{-20}$

$$M = -10 \cdot log_{10}(P)$$
$$M = -10 \cdot log_{10}(10^{-20})$$
$$M = 200$$

Hladina významnosti skóre M v histogramu

V angličtině se tato hodnota udává také jako *significance*. Udává celkovou pravděpodobnost, že shoda je náhodná. Také zahrnuje znalost velikosti databáze, kde se sekvence vyhledává. Tyto informace přináší objektivní měření významnosti výsledků. Nejběžnější hranice akceptování je, pokud se událost vyskytuje s frekvencí menší než 5%. To je hodnota, která je vypsána na hlavní listině výsledků (report).

Hlavní výsledková listina pro typickou peptidovou mapu hmotností udává to, že skóre větší než 67 je významné ($p < 0,05$). Histogram rozložení skóre pro tento případ vypadá takto:

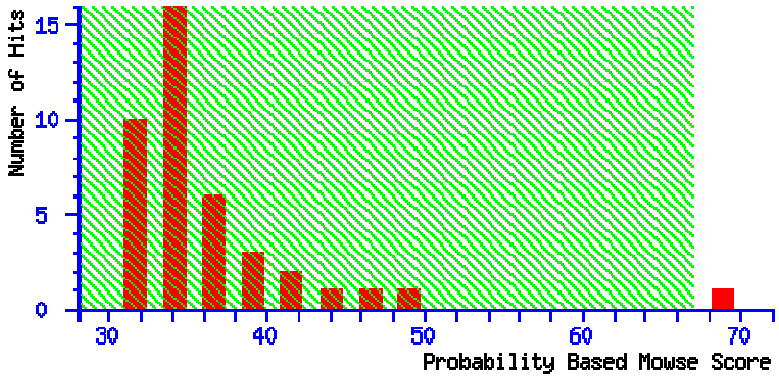


Obrázek 13: *Histogram skóre (červeně) s hladinou významnosti (zeleně) – vysoká kvalita měření. [10]*

V tomto příkladu je analyzovaný protein hodnocený vysokým skóre 108. A je to 26 kDa protein z kvasnic. To je dobrý příklad, protože vysoké skóre je vysoce významné. Skóre v zeleném poli nejsou významné. Skutečná shoda, která nepatří do náhodných jevů, udává skóre, které je mimo oblast.

To je důležité pro rozlišení mezi statisticky významnou shodou a nejlepším výsledkem. V nejlepším případě se významná shoda rovná také nejlepší, s nevyšším skóre. Nicméně významnost je funkce kvality dat. Může nastat situace, kdy nejsou dostačující hodnoty hmotnostního spektra pro statisticky významnou shodu. To však neznamená, že nejlepší shoda (match) není správná. Znamená to ale, že je třeba se více na výsledky zaměřit.

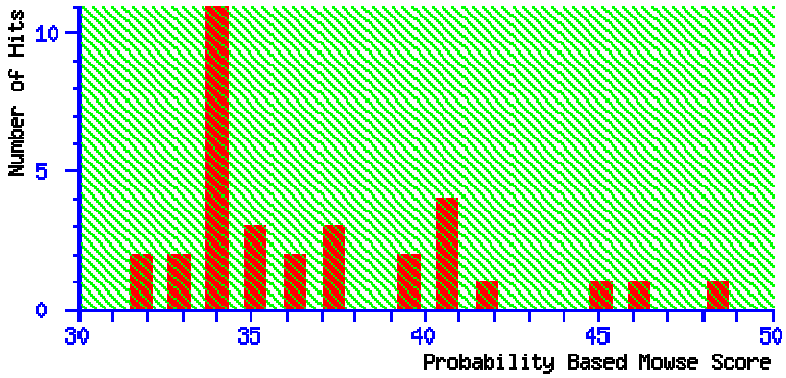
Pro ilustraci rozdílu mezi statisticky významnou a správnou shodou se bude opakovat vyhledávání na příkladu proteinu z kvasnic. Rozdíl bude v navýšení tolerance hmotnosti z ±0,1 Da na ±1,0 Da. Rozlišení vyhledávání se tímto zásahem velmi sníží. Skóre pro správnou shodu selže pro úroveň, kdy začíná mít výsledek statistický význam. Pro ilustraci tuto situaci vyjadřuje graf níže.



Obrázek 14: *Histogram skóre – střední kvalita měření.[10]*

Nejlepší shoda je stále korektní, ale málo významná. Pokud se provede 20 takových vyhledání, může se očekávat shodný výsledek pro nejvyšší hodnotu. Správná shoda je uvedena na prvních pozicích listu správných identifikací (hit list).

Podmínka je udržení se v toleranci hmotností ±2,0 Da. Protože skóre je značně pod rozlišením hranicí významnosti, mohl by být nejistota ve výsledku, pokud je nezjištěno. Nárůst tolerance hmotnosti na ±2,5 Da shody vymizí. Nejvyšší skóre pro náhodné shody je 48.



Obrázek 15: *Histogram skóre - nekvalitní měření.[10]*

V případě neznámého vzorku není přítomen statisticky významný výsledek. V takovém případě je možné prohlásit výsledek za falešně pozitivní.

Tolerance hmotnosti

Hodnota odpovídajících si hmotností je konstantní. Skóre v peptidovém mapování (peptide mass fingerprint) může být inverzně rozděleno do tolerancí hmotností. Tato myšlenka byla využita výše. Tohle však není případ u MS/MS iontového vyhledávání. Zde přírůstek tolerance hmotností peptidů nemá efekt na skóre iontu. To je protože skóre iontů (ions score) vychází z MS/MS fragmentů iontových párů (matches). Otevření tolerance peptidových hmotností znamená, že Mascot má pro testování o mnoho více peptidů. To prodlouží čas srovnávání a redukuje rozlišení. Na iontové skóre to však žádný vliv nemá.

Jestli tolerance hmotností je zvolena příliš pevně, ve snaze zlepšit rozlišení, jeden, nebo více peptidových shod může být ztraceno. To může výrazně ovlivnit celkové skóre.

Omezení

Jako každý statistický přístup, skóre založené na pravděpodobnosti závisí na předpokladech a zjednodušených modelech. Jeden z předpokladů je, že položky v databázi sekvencí jsou náhodné. To není vždy výhodné. Některé případy zahrnují rozšíření opakováním se. Jako příklad je uveden submaxillary apomucin (AAC62527). Ačkoli je molekulární hmotnost tohoto proteinu 1,2 MDa, přes 80% ze sekvence zahrnuje identické 7 kDa opakování (extended repeats). Jestli je jednotlivá (experimentálně získaná) hmotnost peptidu přípustná pro shodu více vypočítaných

hmotností, potom samotná experimentální hmotnost (shoduje se během opakování) nahromadí bezvýznamné skóre. Takové skóre nemá smysl. Pokud nejsou povoleny zdvojnásobené shody, je prakticky nemožné získat shodu pro protein. Pro získání statisticky významných skóre, je hodnota naměřených hmotností příliš malá.

Další hypotéza je, že každé experimentální měření je statisticky nezávislý děj. Nemusí to vždy platit, pokud data zahrnují vícenásobné hodnoty hmotností pro shodný peptid. Stejné to je pro situaci, pokud je vzorek skládající se z iontů s různým nábojem. Správná detekce píků a nastavené úrovně rozhodování jsou nezbytné pro každý algoritmus k získání smysluplného výsledku (skóre).

Postup hodnocení

Aminokyselinové sekvence, nebo složené informace v sobě zahrnují parametry *seq()*, nebo *comp()* a v nich odpovídající hodnoty. Zachází se s nimi jako s filtry na měřené sekvence. Nejednoznačné sekvence, nebo složená data (compositon data) mohou být také použity (situace jako u regulárních výrazů). Stále ale platí jako filtry a ne jako pravděpodobnostní shody typu vyhledávání v Blast nebo Fasta (vyhledávacího algoritmu).

Na druhou stranu, *tag()* a *etag()* parametry jsou hodnoceny jako pravděpodobnosti. To znamená, že je ve skutečnosti více parametrů, než shod s vysokým skóre. Všechny parametry však nejsou potřeba pro shodu. [10]

Nízká hodnota „M“ nemusí vždy znamenat nesouhlas v porovnávání, ale špatnou kvalitu získaných MS/MS dat. Mascot ohodnocení nezohledňuje možnosti využití kombinace komplementárních fragmentačních technik *kolizní aktivační disociace* (CAD) a *disociace elektronovou pastí* (ECD). Tuto možnost zahrnuje práce švédských vědců z university Uppsala. [23]

IDENTITY A HOMOLOGY SKÓRE

Pro upřesnění hodnocení, se základní skóre rozděluje dále na skóre podobnosti (homologie) a totožnosti (identity). Vzorec pro výpočet je stejný, liší se jen zvolenými hodnotami pravděpodobností (tedy prahem).

Identity Skóre

Hodnota shody (skóre identity) vyjadřuje minimální práh skóre. Identity skóre je vypočítáno ze vzorce:

$$M_{ID} = -10\left(log\left(\frac{p}{\#shod}\right)\right) \qquad (4)$$

kde *p* je zvolený práh pravděpodobnosti (pro výpočty se užívá hodnota 1.0), #shod je počet předchozích shod.

To nám dá ekvivalentní ion-skóre (= Mascot skóre „M") jestli je shoda náhodná (náhodným jevem). Jako hlavní předpoklad platí, že teoretický výpočet je velmi konzervativní. Často se stává, že výsledek ion-skóre je pod hranicí identity skóre. Používá se rovnice pro diskriminantní skóre, odečtením identity skóre od ion-

skóre. Základní výpočet se zakládá na hodnotě, o kolik ion-skóre převýší identity-skóre.[24]

Prahová hodnota Identity a Homology

Práh identity (Identity treshold) pro skóre *M* je vypočten z počtu experimentů, například: Pokud je dáno 5000 předchozích shod (= N), je zde pravděpodobnost p = 1:20 = 0,05, že nastane falešně pozitivní shoda. Celková pravděpodobnost *P* je spočtena jako:

$$\boxed{P = p \cdot \frac{1}{N}} \tag{5}$$

$$P = \frac{1}{(20 \cdot 5000)}$$

$$P = 1 \cdot 10^{6}$$

kde odpovídající skóre k pravděpodobnosti je potom:

$$M = -10log_{10}(P)$$
$$M = -10log(1 \cdot 10^{6})$$
$$M = 50$$

Pokud má skóre přiřazenou pravděpodobnost *p*, je potom nastavení prahu významnosti velice jednoduché. Při pravděpodobnosti 1:20 k získání falešně pozitivního výsledku se vyhledává v databázi zahrnující 5000 peptidů, které jsou prekurzory změřených molekulových hmotností. Potom se vyhledává pro celkovou pravděpodobnost menší než $\frac{1}{20 \cdot 5000}$. Pro tuto pravděpodobnost je výsledek prahového Mascot skóre rovno hodnotě 50. Při pravděpodobnosti *p* = 1:200 je prahová hodnota M = 60, 1:2000 je 70, atd.

Ion-skóre, iontové ohodnocení, neboli Mascot skóre „M“. V algoritmu Mascot je skóre pro MS/MS shodu je založeno na absolutní pravděpodobnosti „P“, která vyjadřuje pozorovanou shodu mezi experimentálními hodnotami a daty v databázi jako náhodnou událost. Výsledné skóre je vyjádřeno -10Log této pravděpodobnosti. Pokud při vyhledávání padne $1{,}5.10^{5}$ peptidů do pásma předběžné hmotnosti (precursor mass) a hranice významnosti (kdy má výsledek smysl) je zvolena na 0,05 (pravděpodobnost 1:20 falešně pozitivního výsledku), tak bude práh významnosti skóre (ion-score treshold) na hodnotě 65.

Pokud je špatná kvalita MS/MS spektra, obzvlášť pokud je poměr signál šum nízký, shoda „neporušené“ sekvence nemusí překročit tuto absolutně stanovenou hranici. Z těchto a dalších příčin nabízí Mascot ještě druhý, nižší práh, pro zvýraznění přítomnosti chyb. Nižší, relativní práh je udáván jako „*práh podobnosti*“ (homology treshold). Vyšší, absolutní práh je udáván jako „*práh totožnosti*“ (identity treshold). Jinými slovy: Ideální MS/MS spektrum má jednu, nebo více

kompletních fragmentů iontové série. Žádné nepřiřazené píky šumu a přesně určenou hmotnost. Pokud jsou k dispozici pouze ideální data, není třeba skóre založené na pravděpodobnosti. Reálná data nejsou ideální a nikdy se nezíská dokonalé měření. V případě, kdy shody nejsou perfektní, jsou v zásadě možné tři indikace:

- náhodné shody nemající smysl,
- MS/MS spektrum, které reprezentuje buď sekvence v databázi, nebo jí velmi podobná (significant homology),
- je zde vysoká pravděpodobnost, že MS/MS spektrum představuje přesnou sekvenci v databázi (identitu, nebo značnou podobnost).

Na rozdíl od vyhledávání BLAST, v případě absence ideálních dat se zde nemůže počítat s ideální shodou. Proto se zde preferuje termín „identické, nebo velmi podobné“.

EXPECT

Význam se dá se přeložit jako *věrohodnost vypočteného skóre*. Zahrnuje ve svém výsledku všechny zjištěné informace a jeho výstup má význam pravděpodobnosti správného výsledku (skóre). *Žádaná hodnota je co nejmenší*, tzn. nejvíce věrohodný výsledek.

Hodnoty pravděpodobnosti „p“

Anglicky *expectation values*. Každé skóre proteinu v identifikovaném složení hmotností peptidu a každé skóre iontů v MS/MS vyhledávání je doprovázené pravděpodobnostní hodnotou. Tato hodnota vyjadřuje počet shodných položek s ekvivalentním, nebo lepším skóre, které jsou předpokládané pro výskyt díky pravděpodobnosti. To je ekvivalentní hodnota „E-value“ v BLAST[2] vyhledávacím algoritmu. Více informací o Blast algoritmu je možné najít ve zdroji [25]

U skóre, které je přesně závislé na zvoleném rozhodovacím prahu, je práh defaultně nastaven na ($p < 0,05$). Hodnota pravděpodobnosti (expectation) je potom 0,05. Nárůst skóre o 10 posune hodnotu pravděpodobnosti na 0,005.

Hodnota pravděpodobnosti shody „E_M“ (expectation)

E_M (expect pro Mascot) je hodnota, s jakou pravděpodobností mohu očekávat správný výsledek a vypočítá se vzorce:

$$E_M = P_{prah} \cdot \left(10 \cdot \frac{(M_{prah} - M)}{10}\right) \quad (6)$$

[2] Jeden z nejpoužívanějších prohledávacího nástrojů je Basic Logical Alignment Search Tool. [10]

kde P_{prah} je prahová hodnota celkové pravděpodobnosti, M_{prah} je prahové Mascot skóre odpovídající P_{prah} , také se někdy označuje jako „MIT“ je „Mascot Identity Treshold = práh identity“, M je hodnota skóre, pro kterou se E_M počítá, E_M je pravděpodobnost náhodného (falešně pozitivního) výsledku.

V případě $P_{prah} = 0{,}05$ a $M_{prah} = 50$ jsou výsledky Mascot skóre M a E_M (expectation) následující:

- M = 40, odpovídá E_M=0,5,
- M = 50, odpovídá E_M=0,05,
- M = 60, odpovídá E_M=0,005.

Očekávaná hodnota (expectation) nepřináší žádnou novou informaci. Je odvozeno z výsledků „Mascot skóre“ M a „Prahové hodnoty“ M_{prah}. Výhoda je vyjádření všeho potřebného v jediné hodnotě. **Zcela náhodná shoda má očekávanou hodnotu (expectation) 1, nebo větší. Lepší shoda má hodnotu menší. [**10] **Hodnota expectation (E_M) tedy vyjadřuje míru věrohodnosti skóre.**

Seskupování do proteinových hitů

Report výsledků z vyhledávání zahrnují četné položky. Vždy není jasné, jaký peptid patří ke kterému proteinu. Využití červených a **tučných** řezů písma pro zvýraznění více smysluplných peptidů v proteinech. Nejprve se peptidová shoda vypíše tučně. Vždy při vysokém skóre je navíc položka zvýrazněna červeně. Tzn., že když jsou hity proteinu tučně červené, tak jsou s největší pravděpodobností správně určené. Tyto hity reprezentují nejvyšší skórované proteiny skládající se z jednoho, nebo více vysoce hodnocených peptidových shod.

Shrnutí výsledků pro proteiny

Pro protein je první řádek identifikační řetězec proteinu. Na stejném řádku následuje molekulární hmotnost a skóre pro protein. Pak následuje pravděpodobnost falešně pozitivního výsledku (expect). Potom následuje popis jednotlivých sloupců.

Tabulka 8: *Report nalezených peptidů.*

1. Q9BZX6 **Mass:** 64996 **Score:** 215 **Expect:** 4.2e-016 **Queries matched:** 15

Tripartite motif protein TRIM19 lambda.- Homo sapiens (Human).

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
814.4300	813.4227	813.4708	-0.0481	319	325	0	R.LDAVLQR.I
958.3500	957.3427	957.4079	-0.0652	373	380	0	R.TDGFDEFK.V
1000.3300	999.3227	999.3967	-0.0740	308	315	0	R.DYEEMASR.L
1165.3900	1164.3827	1164.5080	-0.1253	443	452	1	K.MESEEGKEAR.L
1182.4400	1181.4327	1181.5677	-0.1349	34	44	0	R.QPSPSPSPTER.A
1191.5000	1190.4927	1190.6520	-0.1592	161	170	0	K.HEARPLAELR.N
1300.4700	1299.4627	1299.6419	-0.1792	456	467	0	R.SSPEQPRPSTSK.A
1355.5300	1354.5227	1354.6841	-0.1613	361	372	0	R.QEEPQSLQAAVR.T
1423.5200	1422.5127	1422.6779	-0.1652	45	56	0	R.APASEEEFQFLR.C
1426.5700	1425.5627	1425.7364	-0.1737	468	481	0	K.AVSPPHLDGPPSPR.S
1624.7400	1623.7327	1623.8692	-0.1365	359	372	1	R.LRQEEPQSLQAAVR.T
2265.1100	2264.1027	2264.0292	0.0736	526	546	0	R.ELDDSSSESSDLQLEGPSTLR.V
2544.4100	2543.4027	2543.2907	0.1120	547	568	0	R.VLDENLADPQAEDRPLVFFDLK.I
2653.3900	2652.3827	2652.2778	0.1049	482	507	0	R.SPVIGSEVFLPNSNHVASGAGEAEER.V
2882.5000	2881.4927	2881.3777	0.1151	8	33	0	R.SPRPQQDPARPQEPTMPPPETPSEGR.Q

No match to: 1320.4000, 1348.4100, 2550.3000

- **Observed** - Experimentálně získaná hodnota poměru hmotnosti a náboje (m/z)
- **Mr(expt)** - Experimentální hodnota (m/z) transformovaná do relativní molekulové hmotnosti.
- **Mr(calc)** - Relativní molekulová hmotnost vypočtená ze shodných peptidových sekvencí.
- **Delta** - Diference (chyba) mezi změřenou experimentální a vypočtenou hmotností.
- **Start, End** - Celkové pořadí rezidua, začínající od 1 pro N-konečné reziduum celého neporušeného proteinu.
- **Miss** - Množství vynechaných rozštěpených míst.
- **Ion score** - Iontové skóre (není přítomné ve výsledcích peptidových map hmotností u MS/MS).
- **Peptide** - Sekvence peptidu v 1písmenném kódu. Jsou vypsána rezidua, která spojují peptidovou sekvenci v protein.

Celkové skóre pro protein v sobě zahrnuje příspěvky z peptidové mapy (fingerprint), iontové skóre ze všech položek z MS/MS dat a skóre pro jakoukoliv sekvenci kvalifikovanou jako seq, comp, tag a etag. Shrnutí proteinů je korektní report pro vyhledávání z MS peptidových hmotnostních map (peptide mass fingerprint). Není ale korektní výpis pro MS/MS vyhledávání. Konkrétní sada dat z MS/MS zahrnuje peptidové formy vytvářející směs proteinu.

Pokud se zvolí Shrnutí pro proteiny na MS/MS data, je nutné si uvědomit, že proteinové skóre a hodnota pravděpodobnosti (expectation) může být matoucí.

Shrnutí peptidů

Tělo reportu shrnující peptidy v sobě zahrnuje vyčtení proteinů, řazených sestupně podle proteinových skór. Počet hitů proteinu může být specifikován, když je vyhledávání předloženo (submitted). Pokud je nastavení na AUTO, všechny významné shody budou zobrazeny.

Tabulka 9: *Report nalezených proteinů.*

1. A32800 **Mass:** 61016 **Score:** 1195 **Queries matched:** 31
chaperonin GroEL precursor - human
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑ 11	417.1822	832.3498	832.3827	-0.0329	0	45	0.067	1	K.APGFGDNR.K
☑ 12	422.7433	843.4720	843.5065	-0.0345	0	46	0.072	1	K.VGEVIVTK.D
13	430.7328	859.4510	859.4837	-0.0327	0	12	1.7e+002	2	K.IPAMTIAK.N + Oxidation (M)
☑ 15	451.2499	900.4853	900.5280	-0.0427	0	52	0.017	1	K.LSDGVAVLK.V
☑ 16	456.7806	911.5467	911.5803	-0.0337	0	59	0.0028	1	K.VGLQVVAVK.A
☑ 21	480.7447	959.4748	959.5036	-0.0288	0	45	0.07	1	R.VIDALNATR.A
☑ 24	595.7855	1189.5565	1189.6012	-0.0447	0	57	0.0045	1	K.EIGNIISDAMK.K
☑ 25	603.7720	1205.5294	1205.5961	-0.0668	0	(50)	0.018	1	K.EIGNIISDAMK.K + Oxidation (M)
☑ 26	608.3099	1214.6052	1214.6506	-0.0454	0	73	9.7e-005	1	K.NAGVEGSLIVEK.I
☑ 27	617.2857	1232.5569	1232.5884	-0.0315	0	81	1.8e-005	1	K.VGGTSDVEVNEK.K

Pro každý protein je první řádek zahrnuje přístupový kód (řetězec znaků). Potom následuje molekulová hmotnost, potom Skóre bez pravděpodobnostního

základu odvozené z iontového skóru (ion-score). Nakonec nalezených shodných peptidů. Pokud počet položek přesáhne 100, aproximuje se relativní množství pro protein hodnotou zvanou emPAI. Na druhém řádku následuje popis, který shrnuje objevené peptidy. Tabulka zahrnuje sloupce:

- ☑ – Pokud report nezahrnuje přehledovou tabulku, potom rámeček k zatrhnutí pro zvolení položky (queries) pro opakované vyhledání se objeví v prvním sloupci jakéhokoliv řádku zahrnující první výskyt nejvíce hodnocené shody.
- **Query –** Číslo dotazu (vysvětleno níže).
- **Observed** – Experimentální hodnota (m/z).
- **M_r (expt)** - Experimentální hodnota (m/z) transformovaná do relativních hmotností.
- **M_r (calc)** – Vypočtená relativní hmotnost nalezeného peptidu.
- **Delta** – Chybový rozdíl výpočtů mezi experimentální a vypočtenou hmotností.
- **Miss** – Počet chybějících míst enzymu.
- **Ions score** – Pokud jsou zde zdvojené shody stejného peptidu, potom nižší skór shody je uveden *v závorce*.
- **Expect –** Hodnota pravděpodobnosti (expecation value) pro shodu peptidu.
- **Rank –** Pořadí iontové shody (od 1 do 10, kde 1 znamená nejlepší shodu a zvýrazňuje se červeně).
- **Peptide –** Sekvence peptidu v 1-písmenném kódu. Rezidua spojující peptidové sekvence v protein jsou vypsány také.

Ve zkratce následuje pod-menu pro každý protein, který zahrnuje sadu peptidových shod a informace o nich. Také je možné vyvolat, kliknutím na číslo dotazu (query), kompletní seznam shod peptidu pro danou položku (query).

Tabulka 10: *Seznam peptidů ve všech proteinech položky.*

Top scoring peptide matches to query 27
Score greater than 46 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
80.6	-0.03	1+	A32800	K.VGGTSDVEVNEK.K
41.7	-0.03	4	AAQ23524	VGGSSEVEVNEK
24.3	-0.10			R.VGGMVVEGSVKR.D
19.8	-0.02			K.VGGTSMQNHFR.K
17.3	-0.09			K.VNTSTRIGTER.I
16.8	-0.04			R.VGGMVMPGSVQR.D
16.7	-0.12			R.VNTLQGTPIYK.L
16.0	-0.12			K.KDASMVIQTLK.G
14.3	-0.16			R.VGASEAALFLKK.S
13.5	-0.15			R.VNSTTTRTLIK.A

Pop-up vyvolané okno zobrazuje název položky (pokud je přítomna), potom následuje jeden, nebo dva významné prahy – totožnosti (ohraničují identitu) a podobnosti (určují homologii). Potom následuje tabulka zahrnující informace nejvíce skórovaných peptidových shod pro daný dotaz na peptid (query).

- **Score** - Iontové skóre, M skóre.
- **Delta** – Diference (chyba) mezi naměřenou a vypočtenou hodnotou hmotnosti.

- **Hit** – Množství shod (prvního) proteinu zahrnující shody peptidu. Kladné znaménko indikuje, že četné proteiny obsahují shody pro tento peptid.
- **Protein** – identifikační řetězec pro přístup k obsahu shody. Pokud jsou shody k aktuálnímu dotazu peptidu ve více proteinech, potom je kompletní seznam zobrazen v browseru – stavovém řádku (povoluje mezery).
- **Peptide** - Sekvence v 1–znakovém kódu. Pokud jsou použity různé modifikace pro získání shody, modifikované reziduum je podtržené. Pokud jsou rezidua spojující peptidové sekvence stejné ve všech proteinech (v kterých se nachází). Potom tyto rezidua jsou vypsány také, omezené jsou periodami. Pokud peptidová forma protein ukončuje, potom se na posledním místě objeví pomlčka. [10]

Tabulka 11: *Hlavička XML výpisu kódu ze souboru <7mix2.mascot.xml>.*

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl"
href="http://localhost/pepXML_std.xsl"?>
<msms_pipeline_analysis date="2008-02-26T15:18:31"
xmlns="http://regis-web.systemsbiology.net/pepXML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://regis-web.systemsbiology.net/pepXML
http://localhost/pepXML_v18.xsd" summary_xml="7mix2.xml">
  <msms_run_summary
base_name="C:\Proline\data\workdir\2008.2.26_14.51.28.816000\Mascot2
XML3\7mix2" raw_data_type="raw" raw_data=".mzXML">
<sample_enzyme name="trypsin">
<specificity cut="KR" no_cut="P" sense="C"/>
</sample_enzyme>
<search_summary
base_name="C:\Proline\data\workdir\2008.2.26_14.51.28.816000\Mascot2
XML3\7mix2" search_engine="MASCOT"
precursor_mass_type="monoisotopic" fragment_mass_type="monoisotopic"
out_data_type="out" out_data=".tgz" search_id="1">
<search_database local_path="IPI_Human.fasta" type="AA"/>
<enzymatic_search_constraint enzyme="trypsin"
max_num_internal_cleavages="1" min_number_termini="2"/> …
```

2.11.2 Datový formát

Datový formát (Data File Format) je ve formě prostého textu (ASCII) zahrnující seznam informací o jednotlivých naměřených hodnotách, parametry vyhledávání a ostatní volby.
Pro peptidovou mapu (mass fingerprint) může soubor dat zahrnovat seznam hodnot hmotností a velikost intenzity. Ty jsou reprezentovány výškou špiček píků. Mascot automaticky rozpozná následující formáty:

- (*. PKM) – Applied Biosystems Data explorer
- (*. *) – Bruker Analysis AutoXecute Data Report
- (*. XML) – Bruker
- (*. XML) – mzData

Pro MS/MS iontové vyhledávání, musí soubor naměřených dat zahrnovat jednu, nebo více MS/MS hodnot v *Mascot generic formatu (*.mgf)*. Každý MS/MS *dataset je výčet párů hmotnosti a její intenzity*. Dataset je omezený položkou prvního a posledního iontu. Následující formáty jsou pro MS/MS data jsou také podporovány:

Tabulka 12: *Podporované formáty dat kromě MGF*

Název formátu	**Přípona datového souboru**
Finnigan	**(*. ASC)**
Micromass	**(*. PKL)**
Sequest	**(*. DTA)**
PerSeptive	**(*. PKS)**
Sciex API III	**(*. *)**
Bruker	**(*. XML)**
mzData	**(*. XML)**

Do souboru s daty může zahrnovat začleněné vyhledávací parametry. Většina začleněných parametrů se může objevit ve hlavičce data souboru. V Mascot generic formát souboru MGF se může několik parametrů objevit uvnitř MS/MS sadě dat.

Pokud se objeví konflikt mezi hodnotou začleněného parametru a hodnoty vložené do vyhledávání polem ve formuláři, začleněný parametr získá prioritu. Vyhledávání z polí je v základním nastavení pro chybějící hodnoty z datového souboru.

Následující odstavce ilustrují formáty dat v příkladech. Pravidla, která Mascot používá při parsování z datových formátů popisují co je a co není akceptovatelné.

Hierarchický Formát Macotu
V originále *Mascot Generic Format* (*.mgf). Tento formát dat pro navrhovanou (submit) analýzu je obecně popsán takto: (hranaté závorky indikují nepovinné údaje)

Tabulka 13: *Formát dat MGF*.

[Uložené Parametry] Dotaz 1 [Dotaz 2] . . . [Dotaz N]

Pro přehlednost, mohou být mezery použity kdekoliv. Komentář začíná jedním ze symbolů uvedených v závorce (**# ; ! /).** Může však být zahrnut, pouze pokud je v datasetu určen POČÁTEČNÍ ION**T** a KONCOVÝ ION**T**. Ty ohraničují exaktně MS/MS sadu dat.

Peptidová mapa (Peptide Mass Fingerprint)

V případě peptidových map je každá položka samostatná hodnota hmotnosti peptidu. Ta obsahuje dodatečnou hodnotu pro intensitu píku (m/z). Například:

Tabulka 14: *Výpis kódu – možnost popisu píku.*

```
764.2
1231.0
1284
1944.8
2020.2
2100.35
```

,nebo

Tabulka 15: *Výpis kódu – možnost popisu píku.*

```
764.2 2010
1231.0 2345
1284 456
1944.8 1012
2020.2 23
2100.35 566
```

Jestli se vyskytuje v naměřených MS datech (výstupu měření) hodnota na každém řádku, bude ignorována.

Změna vyhledávacích parametrů

Jsou dvě možnosti pro změnu implicitně nastavených vyhledávacích parametrů. První je využití pole formuláře. Další je uložení začleněných (embedded) parametrů na začátek datového souboru. Například:

Tabulka 16: *Implicitní nastavení.*

```
COM=Digest #A6345
CLE=Lys-C
CHARGE=1+
PFA=1

764.2 2010
1231.0 2345
1284 456
1944.8 1012
2020.2 23
2100.35 566
```

Uložené parametry (COM, CLE, CHARGE, PFA) mají nejvyšší prioritu vstupu. Všechny ostatní parametry platí jako defaultně nastavené.

Datový soubor peptidové identifikace hmotností může také zahrnovat pouze vlastní dotazy. Sled dotazů, nebo MS/MS sada dat nejsou dovoleny.

MS/MS iontové vyhledání

Pro MS/MS tandemovou hmotnostní spektrometrii, každý dotaz representuje kompletní MS/MS spektrum. Je určen dvojicí příkazů BEGIN IONS a END IONS.

Vyhledávací formulář je implicitně nastaven a jednotlivé hodnoty mohou být přepsány vloženými parametry na začátku datového souboru. Parametry ve vyhledávacím formuláři, nebo hlavička datového souboru se použije pro celé vyhledávání. Během každého MS/MS dotazu, očekávaná hodnota (prekurzor) hmotnosti peptidu *musí* být určen PEPMASS parametrem. Vložené parametry jsou s každým dotazem volitelné a mohou být použity k určení následujících vlastností:

- TITLE - Názvu hlavičky pro identifikaci spektra.
- CHARGE – Hodnotu náboje předcházejícího peptidu.
- TOL – Tolerance pro peptid.
- TOLU – Svazek tolerancí peptidu.
- SEQ – Dotaz na sekvenci (násobné SEQ dotazy jsou povoleny).
- COMP – Sestavení označení.
- TAG – Označení sekvence (násobné značky TAG jsou povoleny).
- ETAG – Sekvence odolné proti chybám (násobné ETAG parametry jsou povoleny).
- SCANS – Hodnota skenu, nebo rozsahu.
- RTINSECONDS – Doba pro uchování v paměti, nebo časového rozsahu (v sekundách).
- INSTRUMENT – Spárovaná iontová série.
- IT_MODS – Volitelné modifikace.

Parametry uvnitř MS/MS dotazu se použijí jenom lokálně, na jedno spektrum. V případě CHARGE parametru, je myšleno, že je dáno globální CHARGE nastavení. Přesněji konkrétní parametry z vyhledávacího formuláře, nebo hlavičky datového souboru pro lokální nastavení nebo pro více MS/MS dotazů. To může být užitečné, pokud hmotnostní spektrometrický systém dat nemůže vždy určit předchozí stav nabití iontu správně. Například globální nastavení je 2+ a 3+. Pokud bude nejednoznačně definován stav nabití, korektní náboj je zapsán do lokálního CHARGE parametru.

Parametry uvnitř MS/MS dotazu musí být vždy umístěny na začátku, jakmile se objeví značka BEGIN IONS (počáteční iont). Značky se nemůžou objevit uvnitř, nebo v příštím seznamu iontu (ion list). Například:

Tabulka 17: *Výpis souboru MGF.*

```
COM=10 pmol digest of Sample X15
ITOL=1
ITOLU=Da
MODS=Carbamidomethyl (C)
IT_MODS=Oxidation (M)
MASS=Monoisotopic
USERNAME=Lou Scene
USEREMAIL=leu@altered-state.edu
CHARGE=2+ and 3+
```

Tabulka 18: *Výpis souboru MGF.*

```
BEGIN IONS
TITLE=Spectrum 1
PEPMASS=983.6
846.60 73
846.80 44
847.60 67
.
.
.
1640.10 291
1640.60 54
1895.50 49
END IONS

BEGIN IONS
TITLE=Spectrum 2
PEPMASS=1084.9
SCANS=3
RTINSECONDS=25
345.10 237
370.20 128
460.20 108
.
.
.
1673.30 1007
1674.00 974
1675.30 79
END IONS

BEGIN IONS
TITLE=Spectrum 3
PEPMASS=1244.7
SCANS=4-5,7,29-34
RTINSECONDS=26-27,40,95-97
.
.
.
```

Většina MS datových systémů má integrovaný formulář ASCII pro export. Ten je generován do souboru a je možné ho editovat běžným textovým editorem.

Fragmenty informací iontových intenzit jsou velmi důležité. Mascot iterativně volí sub-sadu nejvíce intenzivních hodnot (píků), pro hledání skupiny (group), která nejvíce rozliší skóre pro nejvíce shodný protein. Předcházející hodnota intenzity (prekursor) může být specifikována zahrnutím druhotné hodnoty na PEPMASS řádku parametru. Ten je určen mezerou.

Pro MS/MS iontovou analýzu (MS/MS ion search) je možné do datového souboru v Mascot generic formátu zahrnout sekvenční dotaz a dotazy na peptidovou hmotnostní identifikaci. Není to však možné, pokud soubor obsahuje proprietální (patentovaný) formát MS/MS dat. Jako příklad je uveden výpis ze souboru:

Tabulka 19: *Výpis kódu – proprietální formát.*

```
# following lines define parameters.
# NB no spaces allowed on either side of the = symbol
COM=My favourite protein has been eaten by an enzyme
CLE=Trypsin
CHARGE=2+
# following line will be treated as a peptide mass
1024.6
# following line is a sequence query, which must
# conform precisely to sequence query syntax rules
2321 seq(n-ACTL) comp(2[C])
# so is this
1896 ions(345.6:24.7,347.8:45.4, ... ,1024.7:18.7)
# An MS/MS ions query is delimited by the tags
# BEGIN IONS and END IONS. Space(s)
# are used to separate mass and intensity values
BEGIN IONS
TITLE=The first peptide - dodgy peak detection, so extra wide
tolerance
PEPMASS=896.05 25674.3
CHARGE=3+
TOL=3
TOLU=Da
SEQ=n-AC[DHK]
COMP=2[H]0[M]3[DE]*[K]
240.1 3
242.1 12
245.2 32
.
.
.
1623.7 55
1624.7 23
END IONS
```

Pravidla datových souborů (shrnutí kapitoly „Formát dat pro Mascot“)

- Přípony v názvu souboru nejsou důležité.
- Numerické hodnoty musí být nelokalizované US-ASCII znaky.
- Parametry štítků nejsou citlivé na malé, nebo velké znaky (case sensitive).
- Parametry hlaviček datových formátů se aplikují na vstupu pro vyhledávání a přepisují defaultní nastavení nabízené ve vyhledávacích formulářích.
- Při absenci parametru FORMAT, je jako default nastaven Mascot generic (MGF).
- Mascot generic format povoluje k MS/MS vyhledávání zahrnout dotazy na identifikaci hmotností peptidu a sekvenční dotazy.
- V MGF formátu je každé MS/MS spektrum určeno značkou pro počáteční iont BEGIN IONS a závěrečný iont END IONS. Zde je řádek pro každý fragment píku, určující (m/z) molekulární atomovou hmotnost a intenzitu nabití. Oddělení hodnot je zvoleno mezerou (white space). Fragment m/z hodnot musí být

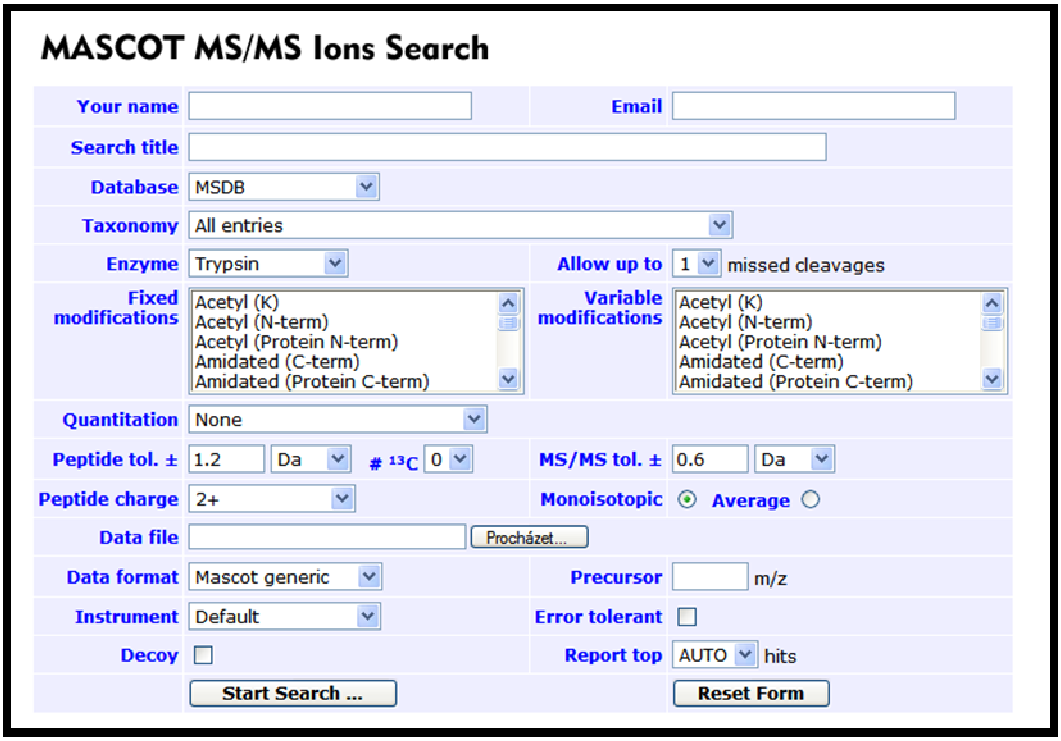
pozitivní a nenulová hodnota. Intenzity musí být pozitivní hodnoty. Jakákoliv další přidaná hodnota nebo text jsou ignorované.

- Parametry mezi BEGIN IONS a END IONS značkami platí jen pro lokální MS/MS dotaz pro jedno spektrum. Výjimku tvoří značky PEPMASS, TITLE, SCANS a RTINSECONDS, které se mohou objevit uvnitř MS/MS bloku dotazu, nebo jako kandidáti pro celkové hodnoty pro použití syntaxe sekvenčních dotazů (Sequence Query syntax).
- Prázdné řádky mohou být použity kdekoliv pro zvýšení přehlednosti.
- Řádky, začínající jedním ze symbolů uvedených v závorce (`# ; ! /`) jsou komentáře a při parsování dat ignorovány. Komentáře nemohou být použity mezi BEGIN IONS a END IONS údaji určujícími MS/MS dotazovací blok.
- Typ vyhledávání SEARCH musí být definován (PMF, SQ, nebo MIS). Implicitně je určeno vyhledávacím formulářem použitým pro soubor po načtení. Jako jakýkoliv jiný parametr, tento může být přepsán zahrnutím SEARCH parametru ve hlavičce souboru.
- Identifikace pomocí mapy hmotností peptidu (peptide mass figerprint, zkráceně PMF) může obsluhovat pouze PMF dotazy. To dovoluje uvolnit více syntaxi, ve které každý řádek začíná hodnotou přiřazené (m/z) významu. Ekvivalentní rozsahu $100 \leq M_r \leq 16\,000$ (kde M_r je relativní hmotnost peptidu).
- MS/MS identifikace může zahrnovat MS/MS data v proprietálním formátu, jenom pokud je deklarován v parametru FORMAT. Mixování proprietálních formátů, nebo zahrnutí jiných, než MS/MS dotazů v proprietálním formátu souboru nejsou povoleny ani podporovány.
- Uživatelské parametry jsou jakékoliv parametry s názvem „USER\d\d", (kde "\d" je digit) nebo jakékoliv jméno začínající s podtržítkem z následujícího seznamu: _INTEGRA_* _DAEMON_* _DISTILLER_* _SERVER_*. Uživatelské parametry nemohou být použity mezi tagy BEGIN IONS a END IONS určující blok MS/MS.

Formát zadání vstupních dat (*on-line)*

Na stránkách firmy Matrix Science je možné provést analýzu Mascot s požadovanými volitelnými parametry. Je dobré respektovat přednastavené hodnoty. Výběrem nevhodných parametrů je možné celou analýzu znehodnotit.

Pro *off-line* analýzu je postup zadávání shodný. Toto prostředí disponuje stejným grafickým prostředím jako je na webu. Z technického hlediska je zde rozdíl v umístění aplikační logiky, která je na počítači (nebo vnitřní síti) s právě puštěným nástrojem Mascot.



Obrázek 16: *Vstupní menu pro analýzu Mascot přeš www rozhranní.*

2.11.3 Nástroj TPP pro Mascot

V anglickém originále *Trans Proteomic Pipeline* (TPP), neboli nástroj pro proteomická propojení dat. TPP je tedy soubor integrovaných nástrojů (tools) pro MS/MS proteomiku. Podrobnější popis TPP a další možné nástroje se dají najít ve zdroji [26] ke stažení.

mzXML je zdroj testovacích dat z databáze [19]. Pracuje se pouze s IPI_HUMAN databazí, nemusi se najit identifikace, pokud se vyhledává proti kategoricky odlišné skupině vzorů.

Mascot neumí přímo pracovat z mzXML formátem, proto je potřeba konvertovat do tzv. *Mascot Generic File* tedy mgf formátu.

Zavolá se v příkazové CON řádce:

```
C:\> mzXML2Search.exe -mgf 7mix.mzXML
```

Pracovní postup přípravy datového souboru:

1. TPP pro Mascot konverze: mzXML ⇨ mgf (mzXML2search.exe).
2. TPP pro Matlab konverze: mgf (Mascot dat) ⇨ xml (Mascot2xml.exe).

2.12 X!Tandem

Projekt organizace *The Global Proteome Machine Organization* nazvaný X!Tandem se zabývá algoritmem pro identifikaci proteinů ze získaných dat metodou MS/MS. X!Tandem je volně šiřitelný jako tzv. *open-source***.** V tomto je veliký rozdíl od Mascotu, který je jako vyhledávací nástroj (engine) *closed-source*. Identifikace v X!Tandemu obecně spočívá ve srovnání získaných spekter (naměřená data) s idealizovanými spektry peptidů v databázi (vzorová data). Nakonec se výsledku srovnání přiřadí hodnota (skóre) stupně příslušnosti k předpokládanému rozpoznání peptidu. Podrobnější informace k algoritmu jsou ve zdroji [27].

Pracovní postup (workflow)

Na začátku (před vyhledáváním) jsou sekvence proteinů rozbaleny do *peptidových listů*. Pro každý peptid (resp. jeho ionty) je zjištěná (vypočtená) hmotnost každého jeho fragmentu (iontu). Ty potom vytvářejí sestrojené ***charakteristické spektrum hmotností***. To se také nazývá tzv. ***mass fingerprint***, nebo ***peptidová mapa*** konkrétního měřeného peptidu. Peptidová mapa je srovnávána s každým naměřeným MS/MS spektrem a odhodnoceno využitím základního nebo rozšířeného schématu pro hodnocení.

X!Tandem rozděluje srovnávací proces do dvou sekvenčních kroků. V prvním jsou zpřístupněny kompletní databáze spekter proteinů umožňující rychlé vyloučení neshodných sekvencí a určení sady proteinů jako možných kandidátů. V druhém kroku se na kandidátech provádí tzv. čisté hledání a identifikace. [28]

Tabulka 20: *Sekvenční workflow X!Tandem.*

Pracovní tok algoritmu (workflow) se dá pro přehled shrnout do posloupnosti:

- Identifikace proteinu z jednoho nebo více peptidů.
- Vytvoření databáze jen rozpoznaných proteinů.
- Rozšířené hledání modifikace peptidů v identifikovaných proteinech z databáze.

X!Tandem, stejně jako Mascot porovnává každé spektrum se všemi možnými kandidáty peptidů v databázi proteinů. Jedna z výhod X!Tandemu je automatické vyhledávání i již modifikovaných peptidů.

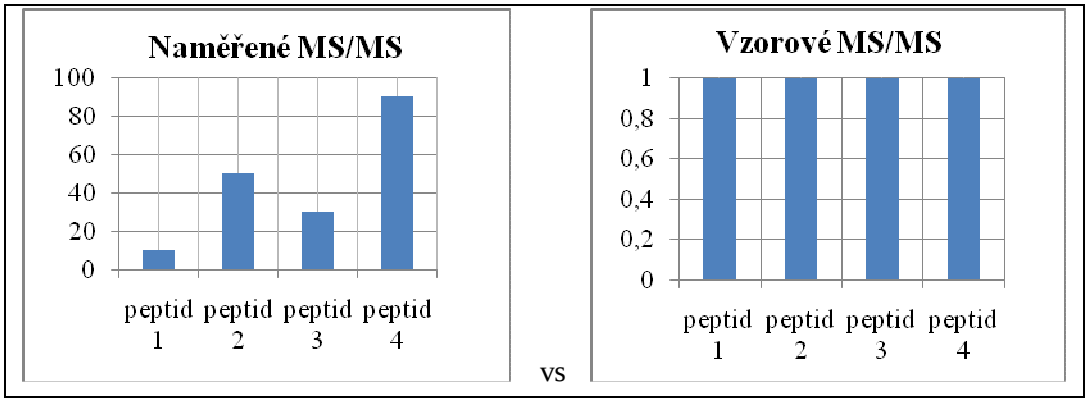
Výhody a nevýhody

Výhody tohoto přístupu spočívají v jeho rychlosti. Je v násobcích přibližně 200x rychlejší pro blíže neurčené hledání a cca. 1000x pro specifické. Příkladem specifického hledání je při hledání oxidace. Další výhody spočívají ve schopnosti

pracovat i se semi-tryptidovými peptidy a polymorfními sekvencemi. X!Tandem také využívá ohodnocení (skóre) založené na pravděpodobnosti odhadu správného výsledku.[29]

2.12.1 Schéma skórování

X!Tandem porovnává naměřené MS/MS spektrum s modelovým MS/MS spektrem. Vzorová spektra jsou založena na peptidech v databázi proteinů. Vzorové spektrum je jednoduše založeno na presenci nebo absenci píku iontů (konkrétně pro ionty typu „y“ a „b“).



Obrázek 17: *Srovnávání hmotnostních spekter (map) peptidů.*

Po sjednocení obou spekter se pro další analýzu použijí jen překrývající se píky. Ostatní (absentující) se pro další zpracování nepoužijí a zahodí. Matematicky se situace realizuje u presentujících vynásobením jedničkou a u absence shody vynásobení nulou.

Předběžné hodnocení $S_{y/b}$ (skóre) je výsledek skalárního součinu (produkt) změřeného a modelového spektra. Po filtraci všech srovnávaných spekter se intenzity překrývajících píků iontů sumují.

$$S_{y/b} = \sum_{i=0}^{n} I_i \cdot P_i \qquad (7)$$

kde „$S_{y/b}$“ je předběžné skóre, „I_i“ je intenzita naměřeného iontu reprezentovaným píkem v grafu, „P_i“ je predikce – absence (=0), nebo presence (=1) vzorového píku.

HYPER-SKÓRE

Předběžné skóre se modifikuje zahrnutím n!, faktoriál počtu „y“ a „b“ měřených iontů. Využití faktoriálu je založeno na hypergeometrické distribuci pravděpodobnosti. Takto upravené hodnocení shody v X!Tandem se označuje jako **Hyperskóre** (hyperscore). Vyjadřuje se písmenem H a spočítá se jako skalární

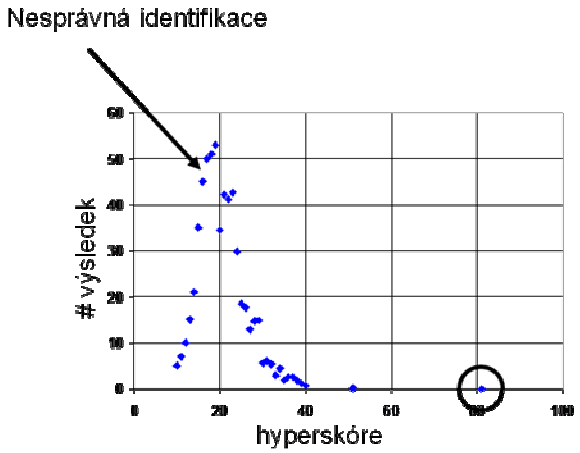
součin faktoriálu[3] počtu nalezených (matched) „y“ a „b“ iontů vynásobený sumou matematicky popsaných píků.

$$\mathrm{H} = N_b! \cdot N_y! \sum_{i=0}^{n} I_i \cdot P_i \qquad (8)$$

kde, „N_b“ je počet „b“ iontů, „N_y“ je počet „y“ iontů, „I_i“ je intenzita (velikost) i-tého píku, „P_i“ je pravděpodobnost výskytu i-tého píku, ta se pohybuje v rozsahu <0,1>.

Grafická interpretace věrohodnosti hyperskóre (expectation value)

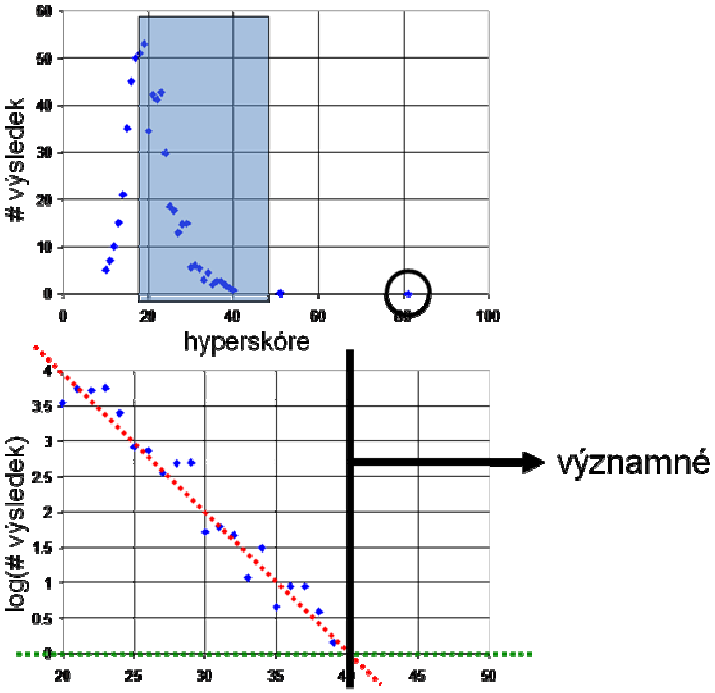
V dalším kroku X!Tandem vytvoří histogram hyperskóre ze všech spekter peptidů, které se nachází v databázi kandidátů. Dále X!Tandem bude pracovat s peptidy s nejvyšším hyperskórem a všechny ostatní s menším hyperskórem se berou za nevýznamné.



Obrázek 18: *Výběr správně určených Hyperskóre.[28]*

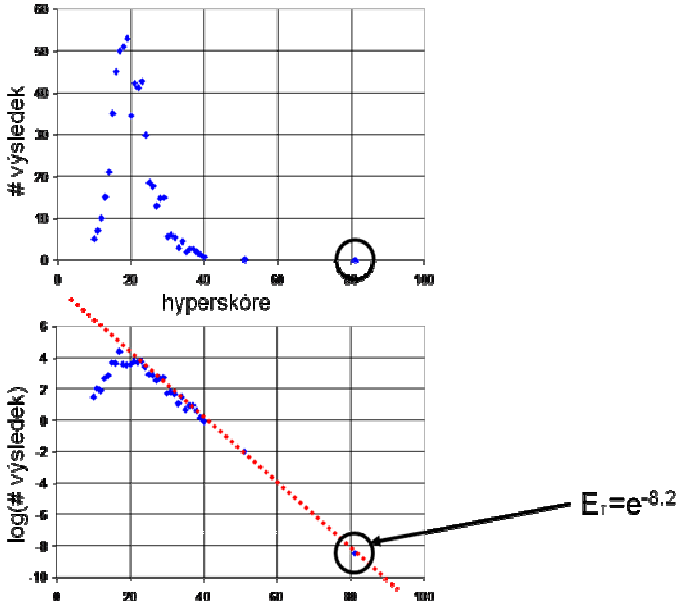
Následně se bere v úvahu druhá polovina histogramu všech dosud spočtených hyperskóre a transformují se na přímku aplikací logaritmu. Nově vzniklá přímka statisticky dokazuje, že nesprávné výsledky jsou náhodné.

[3] Faktoriál (n!) je součin všech „n“ čísel (kladných celých) a vyjadřuje počet způsobů, jak seřadit „n“ různých objektů.



Obrázek 19: *Určení hranice statistické významnosti H.[29]*

X!Tandem tím dokázal, že nejvyšší hyperskóre je jediný správný výsledek nalezení shody (match) a tím určení proteinu. Tato shoda je významná, jestliže je větší než bod, na kterém se protíná přímka s nulovou osou ypsilon a logaritmus výsledků hyperskóre nabývá hodnoty 0. Jakékoliv jiné hyperskóre větší než hranice významnosti jsou nepravděpodobná tomu, že vznikly náhodou.[28]



Obrázek 20: *Graficky určená hodnota E_T. [28]*

EXPECT

Výpočet pravděpodobnostního koeficientu shody (expectation value)

vzorového peptidu (jeho spektrem) s peptidem z naměřených hodnot (jeho spektrem) metodou X!Tandem. Pro každý peptid je spočten histogram hyperskóre ze všech hodnocených spekter. Jen nejvíce hodnocené spektra jsou shodné a validní. Ostatní spektra jsou zahrnuta do náhodných shod. Vyjadřuje pravděpodobnost validního skóre, tzn. pravděpodobnost, že skóre je náhodné a odvozeno z logaritmicko-lineární extrapolace pravé části distribuce hodnot. Násobením této hodnoty počtem hodnocených sekvencí udává hodnotu ohodnocení (expected number). V této hodnotě je zahrnut *peptidový list a daná sada spekter*. Jakmile je určeno peptidové složení, X!Tandem přechází k určení *proteinu*. To je založeno na počtu správně určených peptidů „n“ (hits) v proteinu a jejich partikulárních ohodnocení e_i.

Rovnice je Bayesovský model pro proteiny se shodami o různých hodnotách pravděpodobnosti. První dva vztahy v rovnici popisují pravděpodobnost náhodného původu hmotnostních spekter.

Základem je generování hmotnostního spektra „*s*“. Jestli je proteinová sekvence inferovaná[4] z „*n*“ peptidových sekvencí (jedinečných), každá obdrží hodnotu pravděpodobnosti „e_j“. Vypočtená hodnota pravděpodobnosti *celkové* shody (expectation) pro hledaný *protein* je označena E_T a je vypočítána jako:

$$E_T = \binom{s}{n} \cdot \left(\frac{\beta^n (1-\beta)^{s-n}}{s N^{n-1}}\right) \cdot \left(\prod_{j=1}^{n} e_j\right) \tag{9}$$

$$E_T = \left(\prod_{j=0}^{n-1} \frac{(s-i)}{(n-i)}\right) \cdot \left(\frac{\beta^n (1-\beta)^{s-n}}{s N^{n-1}}\right) \cdot \left(\prod_{j=1}^{n} e_j\right) \tag{10}$$

Kde význam jednotlivých členů rovnice je následující:

- „E_T“ je celková pravděpodobnost shody pro analyzovaný protein.
- „e_j“ je pravděpodobnost shody pro j-tou sekvenci.
- „n“ je celkový počet jedinečných sekvencí.
- „s“ je spektrum molekulových hmotností (pořadí v sadě dat).
- „N“ je počet peptidových sekvencí nalezených peptidů (jedinečných)
- „i“ je index pořadí.
- „j“ je index pořadí.
- „β“ je normalizovaný počet peptidů.

Hodnota β se pohybuje v intervalu <0,1> a je vypočtena ze vztahu:

$$\beta = \frac{N}{\text{celkový počet peptidů v uvažovaném proteomu}} [-] \tag{11}$$

Na celkový výsledek pravděpodobnostní shody „E“ pro protein, mají vliv dílčí hodnoty pravděpodobnosti shody pro jednotlivé peptidy „e_j“, z nichž se protein skládá. Hodnoty peptidů jsou kombinované jednoduchým Bayesovským modelem

[4] Inference - úsudek

pro klasickou pravděpodobnost. Ve speciálním případě, kdy je pozorovaný jen jeden peptid, se vzorec zjednoduší na tvar:

$$E_T = e_1 \tag{12}$$

Hodnota expectation (E_T) tedy vyjadřuje to, jak nepravděpodobné je výsledné hyperskóre pro celý protein (čím nižší, tím lepší) . Čím je hodnota expectation vyšší, tím je větší možnost, že náhodný. X!Tandem spočte E_T extrapolací přímky logaritmu histogramu. Čím vyšší hodnota expectation, tím se dá méně věřit vypočtenému hyperskóre. Žádané jsou tedy proteiny s nízkou hodnotou expectation.[30]

2.12.2 Datový formát

Formát dat formuláře

Algoritmus analýzy X!Tandem pracuje s rozhranním API (apllication programming interface). "X!" komunikuje v jazyce s názvem BIOML[5]. Ten je založený na XML technologii zápisu. [31]

Obecná forma jazyku pro zápis BIOML dokumentu se označuje zkratkou GAML (Generalized Analytical Markup Language).[32]

Zprostředkování komunikace se realizuje *formuláři.* Informace ve všech formulářích jsou kategorizovány do logických skupin.

Vstup a výstup

Vstupem je soubor (formulář) s naměřenými daty. Například soubor s názvem „7mix2.mzXML", který bude dále předmětem analýzy. Výstup je výsledek analýzy opět ve formě datového souboru (formuláře) „7mix2.tandem.xml". Vstup i výstup je validní XML formát, který může být editován v jakémkoliv textovém editoru (např. PSpad[6]).

Výstupní formulář

Zjednodušená struktura vypadá takto:

- PROTEIN - A:
  - **Skupina 1 (group)**
    = ID identifikovaného proteinu s nejvyšší pravděpodobností
    - **Protein 1 (nejlépe hodnocený – nejnižší expect, nejvyšší skóre)**
      - **Peptid proteinu 1:**
        - **Hyperskóre**
        - **Expect**
    - **Protein 2 (jiný)**

[5] *BIOML* je jazyk původně vytvoření pro ukládání informací o biopolymerech.
[6] *PSpad* je freeware program pro editaci post skriptových (textových) souborů. [37]

- **Peptid proteinu 2:**
  - **Hyperskóre**
  - **Expect**

o **Skupina 2 (group)**
= ID identifikovaného proteinu s nejvyšší pravděpodobností

Popis výstupního XML formuláře s výsledky analýzy řádek po řádku.

Tabulka 21: *Hlavička (header)*

TAG	Význam
<?xml version="1.0"?>	Verze XML standardu.
<?xml-stylesheet type="text/xsl" ref="/tandem/tandemstyle.xsl"?>	Definice stylu textu.
<bioml xmlns:GAML="http://www.bioml.com/gaml/" label="test">	Identifikuje dokumentu BIOML.

Skupiny představují souhrn peptidů identifikovaného proteinu. V XML struktuře je realizace formou ploché databáze. Rozlišení je příslušnost k identifikovanému proteinu určujícím ID číslem vzorové databáze, např. IPI.

Tabulka 22: *Popis typů skupin (groups)*

TAG	Význam skupiny
<group> … /<group>	Tělo XML tvořící základ.
<group type="model">	Tato skupina zahrnuje všechny informace o identifikaci jednotlivých peptidů: originální hmotnostní spektrum, histogramy statistik, sekvence souhlasných peptidů a proteinové sekvence zahrnující tyto peptidy.
<group type="parameters">	Parametry pro vyhledávání.
<group type="support">	Histogramy potřebné pro popis výsledků.

Každá skupina se kromě poznámek a dodatečných informací skládá z proteinů. Proteiny určené z peptidů popisují domény.

Tabulka 23: *Příklad zápisu - model skupiny (model groups): <group type="model"> </group>*

Příklad zápisu
<group id="46" mh="1955.165047" z="3" expect="2.4e-005" label="ENSP00000306469" type="model" sumI="3.15" maxI="340.56" fI="3.44">

Tabulka 24: *Význam parametrů modelů skupiny (model groups)*

Parametr	Význam parametru
Id	Hodnota závislá na identifikovaném spektru hmotností.
Mh	Množství iontů (protonů).
Z	Iontový náboj.
Expect	Pravděpodobnost správné identifikace *proteinu* (identifikace změřeným MS).
Label	Text z proteinové sekvence FASTA, souborový popis identifikovaného proteinu (proteinu s nejvyšší pravděpodobností shody).
sum(I)	Log_{10} ze sumy všech fragmentů hodnot napětí iontů.
max(I)	Maximum ze všech změřených částečných intenzit iontů.
f(I)	Multiplikátor (násobitel) pro konverzi do normalizovaného spektra.

Proteiny jsou sekvenčně porovnávány se skupinami „model groups“. Každá skupina má tuto strukturu zápisu:

<protein><note></note><file/><peptide><domain><aa />
</domain></peptide></protein>

Každý model skupiny může zahrnovat mnoho **proteinových elementů**. Každý z elementů obsahuje specifickou informaci. Ta je popsána v tabulce níže.

Tabulka 25: *Přehled kategorií popisek proteinového elementu.*

TAG	Význam
<protein>	Popis parametrů souboru.
<note>	Popisné poznámky o proteinu.
<file>	FASTA název souboru i s cestou.
<peptide>	Peptidové sekvence proteinu.
<domain>	Popis regionu proteinové sekvence, která byla identifikovaná.
<aa>	Popis specifických modifikací .

Detaily každého z parametrů jsou následující:

Tabulka 26: *Parametry kategorie Protein.*

< protein expect="-4.6" id="46.1" uid="30707" label="ENSP00000306469" umI="3.46" >	
Parametr	**Význam**
Expect	Log10 z hodnoty pravděpodobnosti.
Id	Identifikátor lokální identifikace.
Uid	Vygenerované unikátní číslo označení proteinu.
Label	Popiska FASTA souboru.
sum(I)	Součet všech fragmentů iontů, které identifikují protein.

Značka poznámky <note> obsahuje také i užitečné pole pro název identifikovaného proteinu.

Tabulka 27: *Parametry kategorie „Note“.*

<note label="description">IPI:IPI00555997.1|...|VEGA:OTTHUMP00000160895 Tax_Id=9606 Gene_Symbol=MYH2 MYH2 protein</note>	
Parametr	**Význam**
Label	Popiska poznámky.

Tabulka 28: *Parametry kategorie „File“.*

<file type="peptide" URL="../fasta/human_e.fasta.pro"/>	
Parametr	**Význam**
Type	Peptid je jediná platná hodnota.
URL	Cesta k původnímu FASTA souboru.

Tabulka 29: *Parametry kategorie „Peptide“.*

<peptide start="1" end="376">	
Parametr	**Význam**
Start	Hodnota pozice začátku sekvence (peptidu).
End	Hodnota pozice konce sekvence.

Každý s proteinových elementů obsahuje podskupinu oblastí – domén (domain). Ty představují identifikované rámce (peptidy) ze kterých se skládá celkový protein. Jen domény v sobě zahrnují hyperskóre, hodnocení X!Tandemu.

Tabulka 30: *Parametry kategorie Domain.*

```
<domain
id="46.1.1"
start="97"
end="114"
expect="2.4e-005"
mh="1954.064"
delta="1.101"hyperscore="1.101"
peak_count="16"
pre="NELR"post="INRE"
seq="VAPDEHPILLTEAPLNPK"missed_cleavages="0">
```

Tabulka 31: *Parametry kategorie Domain – vysvětlení významu.*

Parametr	Význam
Id	Identifikátor.
Start	První reziduum.
End	Poslední reziduum.
Expect	Pravděpodobnost náhodné shody peptidu. Nejvíce spolehlivé skóre hodnotící nalezený protein. Zahrnuje statistický přístup.
Mh	Vypočtená hmotnost peptidu.
Delta	Změřené MS mínus vypočtené MS.
Hyperscore	Skóre identifikace metodou X!Tandem. „Hyper“ skóre je „raw“ nejlepší skóre bez jakékoliv znalosti o množství hodnocení ostatních peptidů ve spektru výsledků.
Nextscore	Druhé nejlepší skóre. Hodnota vrací rozdíl (delta) mezi nejlepším a druhým nejlepším skórem. Velmi dobrá hodnota rozlišovací schopnosti.
peak_count	Počet píků, které souhlasí mezi teoretickým a testovacím hmotnostním spektrem.
Pre	Čtyři rezidua předchozí domény.
Post	Čtyři rezidua následující domény.
Seq	Sekvence aktuální domény.
missed_cleavages	Počet možných štěpení v aktuální peptidové sekvenci.

Tabulka 32: *Parametry kategorie „aa“.*

<aa type="C" at="247" modified="57.01" />	
Parametr	**Význam**
type	Zkratka pro modifikované reziduum.
at	Číslo rezidua v proterino-peptidových sekvencích.
modified	Změna hmotnosti rezidua zapříčiněné modifikací.

V praxi vypadá příklad analýzy (výstupního formuláře popisovaného vzoru) takto:

Tabulka 33: *Výpis části výstupního formuláře ze souboru <7mix2.tandem.xml>.*

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="tandem-style.xsl"?>
<bioml xmlns:GAML="http://www.bioml.com/gaml/" label="models from
'7mix2.mzXML'">
<group id="732" mh="914.052724" z="2" expect="3.5e-003"
label="IPI:IPI00025879.1|SWISS-
PROT:P12882|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954|H-
INV:HIT000069703|VEGA:OTTHUMP00000160893;OTTHUMP00000183342..."
type="model" sumI="6.80" maxI="664429" fI="6644.29" >
```

Tabulka 34: *Výpis části výstupního formuláře ze souboru <7mix2.tandem.xml> - pokračování.*

```
<protein expect="-166.7" id="732.1" uid="6986"
label="IPI:IPI00025879.1|SWISS-PROT:P12882
|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954
|H-INV:HIT000069703|VEGA:OTTHUMP00000160893;OTTHUMP00000183342..."
sumI="9.00" >
<note label="description">IPI:IPI00025879.1|SWISS-
PROT:P12882|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954|H-
INV:HIT000069703|VEGA:OTTHUMP00000160893;OTTHUMP00000183342
Tax_Id=9606 Gene_Symbol=MYH1 Myosin-1
</note>
<file type="peptide" URL="IPI_Human.fasta"/>
<peptide start="1" end="1939">MSSDSEMAIF …… VNKLRVKSRE VHTKIISEE
<domain id="732.1.1" start="85" end="91" expect="3.5e-003"
mh="913.398" delta="0.655" hyperscore="30.8" nextscore="18.4"
y_score="12.9" y_ions="6" b_score="12.3" b_ions="4" pre="NPPK"
post="AMMT" seq="YDKIEDM" missed_cleavages="6">
</domain>
</peptide>
</protein>
```

Sady skupin s histogramy

Následující proteinové elementy jsou sady jednotlivých skupin zahrnující histogramy s podporou informací relevantních pro identifikaci. Tyto histogramy jsou reprezentovány v GAML verzi. Celkový formát je přerušovaný, se skupinami.

První skupina (supporting data) obsahuje sadu histogramů, které byly vypočítány během identifikačního procesu.

Tabulka 35: *Výpis kódu první skupiny GAML verze.*

```
<group label="supporting data" type="support">
<GAML:trace label="46.hyper" type="hyperscore expectation function">
        </GAML:trace>
<GAML:trace label="46.b" type="b ion histogram">
        </GAML:trace>
<GAML:trace label="46.y" type="y ion histogram">
        </GAML:trace>
</group>
```

Druhá skupina (fragment ion mass spectrum) obsahuje histogramy reprezentující píky v hmotnostním spektru, které byly aktuálně použity vyhledávacím prostředím pro identifikaci.

Tabulka 36: *Výpis kódu druhé skupiny GAML verze*

```
<group type="support" label="fragment ion mass spectrum">
<note label="Description">  </note>
<GAML:trace id="46" label="46.spectrum" type="tandem mass pectrum">
        </GAML:trace>
</group>
```

Histogramy jsou pořízeny ve standardním formátu GAML. Data jsou řazena s „x“ a „y“ koordináty. Koordináty jsou zaznamenány v ASCII znacích. Oddělení znaků je mezerou. Jako příklad je uveden konkrétní záznam:

Tabulka 37: *Výpis kódu – GAML histogram.*

```
<GAML:trace>
<GAML:Xdata ><GAML:values format="ASCII" numvalues="4">
173.465        175.114        177.928        201.136
</GAML:values>
</GAML:Xdata>
<GAML:Ydata> <GAML:values format="ASCII" numvalues="4">
2        7        2        8
</GAML:values>
</GAML:Ydata>
</GAML:trace>
```

Skupina označená *type=“parameters“* zahrnují *informace pro proces identifikace*. Typicky jsou tři na konci souboru (opět ve formě skupin). Parametry jsou stejné jako už popsané ve vyhledávacím prostředí výše.

1. <group label="input parameters" type="parameters">
2. <group label="unused input parameters" type="parameters">
3. <group label="performance parameters" type="parameters">

Podrobnější popis podskupin je následující:

1. Využité vstupní parametry „input parameters“ - Tato skupina zahrnuje všechny parametry použité ve vyhledávacím prostředí.
2. Nevyužité vstupní parametry „unused input parameters“ - Obsahuje všechny parametry zahrnuté do vyhledávání, ale neschopné interpretace pro použití. To nastává ve třech případech:
   - chyby čtení
   - speciální příkazy na rozhranní, které vyhledávací prostředí neakceptuje
   - příkazy, pro jiné verze série vyhledávání „X!“.
3. Výkonové parametry „performance parameters“ - Zde se nacházejí všechny informace získané při vyhledávání. Je to například počet správných identifikací, a počet použitých spekter při srovnávání se vzory v databázi.

Všechny tyto skupiny ve svých záznamech používají tag pro poznámky <notes>. Formát těchto poznámek je následující.

Tabulka 38: *GAML formát poznámek.*

Příklad zápisu TAGu <notes>
<note type="input" label="spectrum, path">t.mgf</note>

Tabulka 39: *Parametry značky „poznámka“.*

Parametr	**Význam**
Type	Specifikuje jaká informace je užita.
Label	Určuje identifikátor popisu.

Data uvnitř, TAGu představují hodnotu pro parametr určený položkou label. V tomto případě jde o „<note>t.mgf</note>“.

Uživatelské prostředí

Je ve více možných provedení. Jednou z možností je off-line režim, který umožňuje spustit analýzu bez přístupu na internet.

Seznam souborů potřebného pro analýzu metodou X!Tandem:

1. databáze vzorů (zvolena IPI v 3.39): **IPI_HUMAN.fasta**
2. zkompilovaný algoritmus X!Tandem: **tandem.exe**

Seznam formulářů potřebných pro analýzu metodou X!Tandem:

3. naměřená MS/MS data (volitelný název): **7mix2.mzXML**
4. implicitní vstupní hodnoty: **default_input.xml**
5. kritéria pro hodnocení: **taxonomy.xml**
6. vstupní soubor se základními parametry: **tandem_in.xml**
7. výstupní soubor analýzy (volitelný název): **7mix2.tandem.xml**

Ve formuláři s názvem „tandem_in.xml“ se zadává název vstupního a název výstupního souboru.

Tabulka 40: *výpis formátu parametrů názvu vstupního a výstupního souboru v „tandem_in.xml“.*

<note type="input" label="spectrum, path">**7mix2.mzXML**</note> <note type="input" label="output, path">**7mix2.tandem.xml**</note>

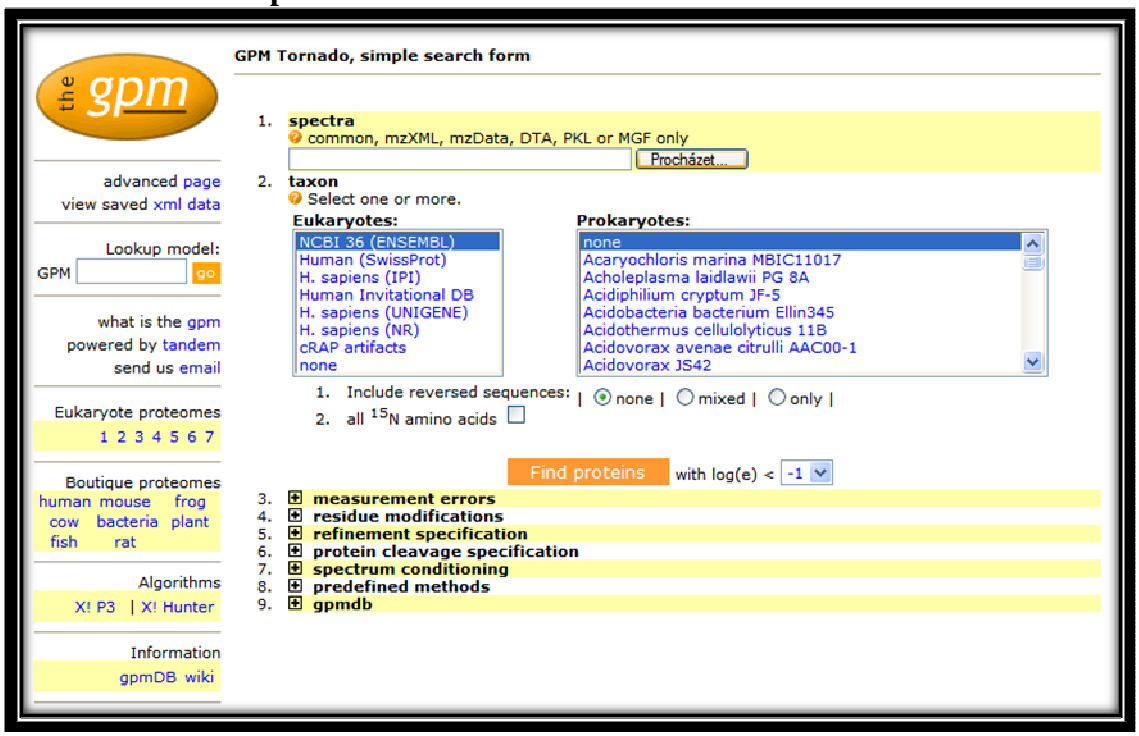
Formát zadání vstupních dat v off-line režimu

1. **vstup** zadaný v příkazovém řádku:
   C:\> tandem.exe tandem_in.xml
2. **výstup** vygenerovaný programem ve stejném adresáři:
   C:\> 7mix2.tandem.xml

Vhodné je vytvoření dávkového souboru (*. bat), do kterého se vstupní příkazy přeprogramují.

Další možností je on-line přístup přes webové rozhranní na domovské www stránce X!Tandemu.

Formát zadání vstupních dat v on-line režimu

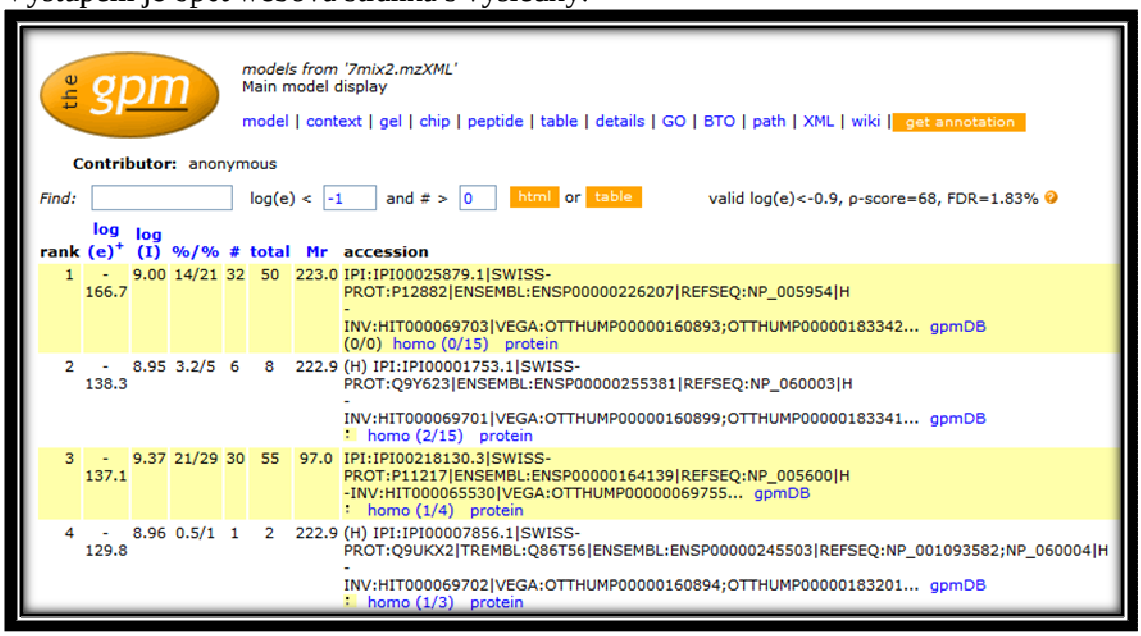


Obrázek 21: *Webová stránka pro zadání vstupních dat.[33]*

Zvolené parametry pro on-line analýzu:

1. Spectra: mzXML soubor naměřených raw dat.
2. Taxon
   - Eukaryotes: H. sapiens (IPI)
   - Prokaryotes: none <default>
3. Mesurement errors: <default>
4. Residue modifications: <default>
5. Refinement specification: <default>
6. Protein cleavage specification: <default>
7. Spectrum conditioning: <default>
8. Predefined methods: <default>
9. Gpmdb: <default>

Výstupem je opět webová stránka s výsledky:



models from '7mix2.mzXML'
Main model display

model | context | gel | chip | peptide | table | details | GO | BTO | path | XML | wiki | get annotation

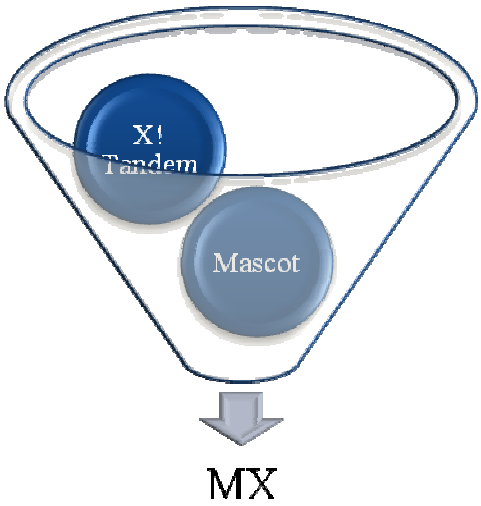
Contributor: anonymous

Find: log(e) < -1 and # > 0 html or table valid log(e)<-0.9, p-score=68, FDR=1.83%

rank	log (e)+	log (I)	%/%	#	total	Mr	accession
1	-166.7	9.00	14/21	32	50	223.0	IPI:IPI00025879.1|SWISS-PROT:P12882|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954|H-INV:HIT000069703|VEGA:OTTHUMP00000160893;OTTHUMP00000183342... gpmDB (0/0) homo (0/15) protein
2	-138.3	8.95	3.2/5	6	8	222.9	(H) IPI:IPI00001753.1|SWISS-PROT:Q9Y623|ENSEMBL:ENSP00000255381|REFSEQ:NP_060003|H-INV:HIT000069701|VEGA:OTTHUMP00000160899;OTTHUMP00000183341... gpmDB homo (2/15) protein
3	-137.1	9.37	21/29	30	55	97.0	IPI:IPI00218130.3|SWISS-PROT:P11217|ENSEMBL:ENSP00000164139|REFSEQ:NP_005600|H-INV:HIT000065530|VEGA:OTTHUMP00000069755... gpmDB homo (1/4) protein
4	-129.8	8.96	0.5/1	1	2	222.9	(H) IPI:IPI00007856.1|SWISS-PROT:Q9UKX2|TREMBL:Q86T56|ENSEMBL:ENSP00000245503|REFSEQ:NP_001093582;NP_060004|H-INV:HIT000069702|VEGA:OTTHUMP00000160894;OTTHUMP00000183201... gpmDB homo (1/3) protein

Obrázek 22: *Webový výstup v příkladu prvních čtyř položek.*

2.13 Metaskóre



Obrázek 23: *Princip vytvoření metaskóre.*

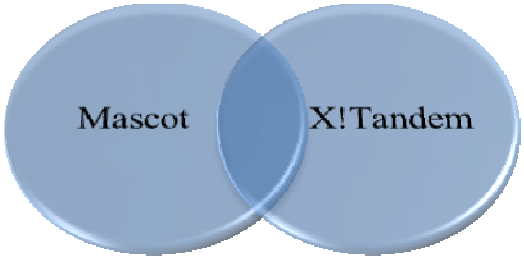
Každá z metod (X!Tandem a Mascot) udává vlastní seznam identifikovaných proteinů společně s vlastními výsledky ohodnocení. Hodnoty hodnocení a seznam identifikovaných proteinů u každé metody liší. To je v obou případech dáno vlastním odlišným přístupem. Smyslem metaskóre je sjednocení výsledků analýzy Mascot a X!Tandem tak, aby uživatel dostal sjednocený výsledek. Tento výsledek je ve formě vlastního seznamu identifikovaných proteinů. Každá položka proteinu je ve výsledku hodnocena jedním určujícím ukazatelem nazvaným Metaskóre a značený zkratkou

MX. Jako doplněk se mohou uvést hodnoty Mascot a X!Tandem výsledků. Obecný tvar seznamu je uveden v tabulce níže.

Tabulka 41: *Výsledek zpracování – seznam identifikací.*

Nalezeno	Ukazatel 1	Ukazatel 2	Metaskóre
Protein A	Mascot	X!Tandem	MX
Protein B	Mascot	X!Tandem	MX
Protein C	Mascot	X!Tandem	MX

Počet nalezených je menší než v každé metodě, protože se zde aplikuje výpočet metaskóre na průnik obou množin dat.



Obrázek 24: *Princip průniku množin vstupních výsledků.*

Výpočet Metaskóre

Rovnice se dá principiálně přirovnat k váženému průměru kde se ke každému proteinu přiřazuje hodnota MX složená ze součtu násobků vah (expect) s ohodnocením (skóre). Součet všech dvojic (obou metod) se vydělí hodnotou, vyjadřující počet těchto párů. Matematická formulace je uvedena níže.

$$MX = \frac{\Sigma_i H_i \cdot E_{T\,i} + \Sigma_j M_j \cdot E_{M\,j}}{N} \quad [-] \qquad (13)$$

Kde z příspěvku X!Tandemu: „H“ je nejlepší Hyper-skóre každé skupiny (group, žádané je co nejvyšší). „E_T“ je nejlepší Expect skupiny (pravděpodobnost špatného výsledku, žádaná co nejnižší). „i“ je index aktuálního pořadí skupiny přiřazené pro nalezený protein. Mascot data přispívají: „M“ je nejlepší Ion-skóre každé skupiny (žádané je co nejvyšší). „E_T“ je nejlepší Expect skupiny (pravděpodobnost špatného výsledku, žádaná hodnota je co nejnižší). „j“ je index aktuálního pořadí skupiny přiřazené pro nalezený protein. „N“ je celkový počet všech položek pro nalezený protein v obou množinách dat (Mascot + X!Tandem).

Do výsledné tabulky se hodnoty vynáší principiálně takto:

Tabulka 42: *Systém zpracování obecně pro tři nalezené a identifikované proteiny.*

Protein	Mascot	X!Tandem	Metaskóre
A	Suma ($skóre_i * expect_i$)	Suma ($skóre_j * expect_j$)	skóre
B	Suma ($skóre_i * expect_i$)	Suma ($skóre_j * expect_j$)	skóre
C	Suma (skóre * expect)	Suma ($skóre_j * expect_j$)	skóre

V X!Tandemu, každá skupina náležící stejnému proteinu obsahuje hitparádu možných výsledků. Po vybrání kandidáta s nejlepším hodnocením si jeho hodnoty dosadí do své preambule (hlavičky). Za platnou hodnotu se bere tedy nejlepší kandidát v dílčí skupině, kterého představuje konkrétní protein. Tento protein je rozpoznaný díky nalezené části svého dílčího řetězce, peptidu. Ostatní kandidáti se zanedbávají.

Tabulka 43: *Dílčí skupina u X!Tandem.*

Skupina 1 = Protein 1	Mascot	X!Tandem	MX
Protein 1 = Peptid ID1	hodnocení	hodnocení	hodnocení
~~Protein 3 = Peptid ID2~~	~~hodnocení~~	~~hodnocení~~	~~hodnocení~~
~~Protein 2 = Peptid ID3~~	~~hodnocení~~	~~hodnocení~~	~~hodnocení~~

2.14 Specializované organizace

Česká společnost pro biochemii a molekulární biologii

Zkráceně ČSBMB. Je pracovištěm Vysoké školy chemicko-technologické v Praze. Jeho proteomická sekce byla založená v roce 2003. Je to občanské sdružení seskupující lidi se společným zájmem o proteomiku.[34]

Ústav hematologie a krevní transfuze

Zkráceně ÚHKT. Provádí specializovanou léčbu, výzkum a vzdělávání. Součástí této organizace je i výzkumný úsek s několika odděleními. Jedno ze zaměření výuky je i oblast proteomiky. [35]

Seznam organizací působící v dané problematice a jejich zkratek se nalezne ve zdroji [36].

2.15 Popis testovacích dat

Vstupem byl mix sedmi proteinů.

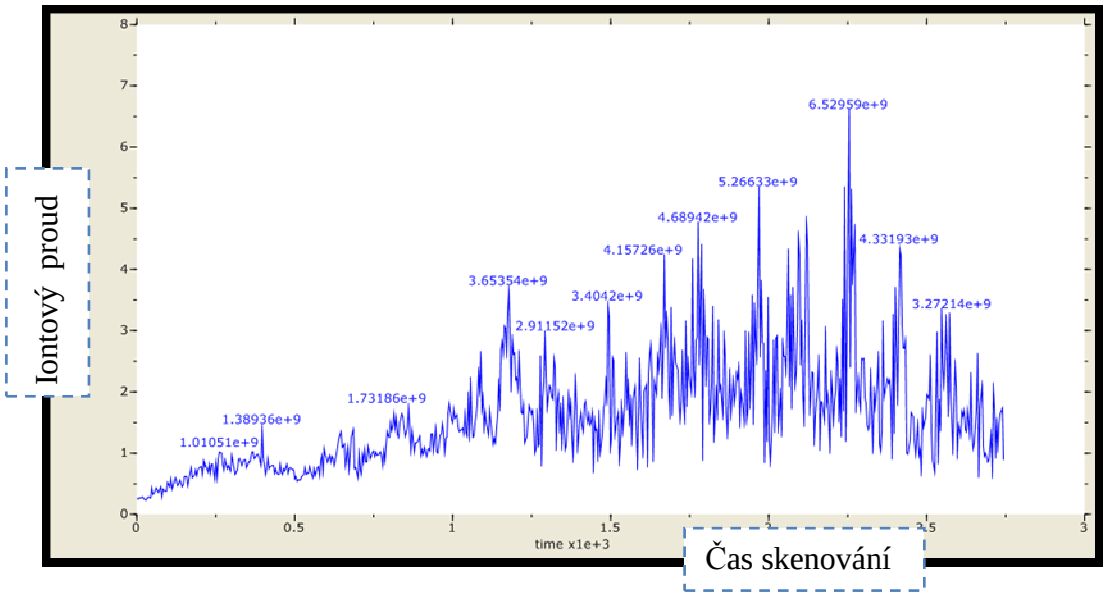
7 protein mix:

1. Rabbit glycogen phosphorylase
2. E. Coli Beta
3. Galactosidase
4. Bovine serum albumin
5. Myosin
6. Chicken Ovalbumin
7. Bovine serotransferrin

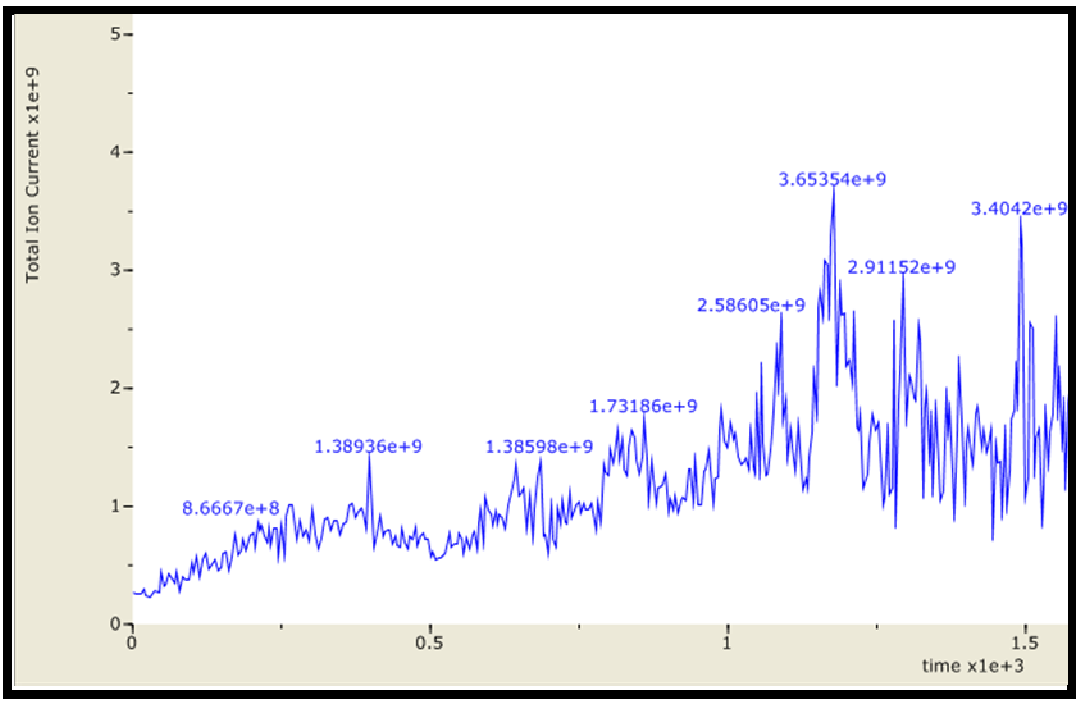
2.15.1 Naměřená MS/MS data

Tento datový soubor byl nahrán z databáze na internetu. Zdroj [19].

Originální soubor nese název „7MIX_STD_110802_1“. Pro testovací účely byl originální databázový soubor zkrácen na první polovinu naměřených dat.



Obrázek 25: *Celkový přehled vzorových vstupních dat souboru „7mix2.xml“.*



Obrázek 26: *Zvětšený začátek naměřených dat souboru „7mix2.xml“.*

2.15.2 Výstup z Mascotu

Červeně vyznačené hodnoty indikují nejlépe hodnocené shody peptidu pro dané spektrum. Nemusí to ale znamenat významnou shodu (significant match). Tučně vyznačené hodnoty vyznačují prvotní shodu pro aktuální spektrum v reportu. Pokud shoda proteinu (hit) neobsahuje tučně červeně vyznačené shody nebo pokud se vyskytují opakované shody, jsou výsledky přiřazeny lépe hodnocenému proteinu nebo proteinům.[10]

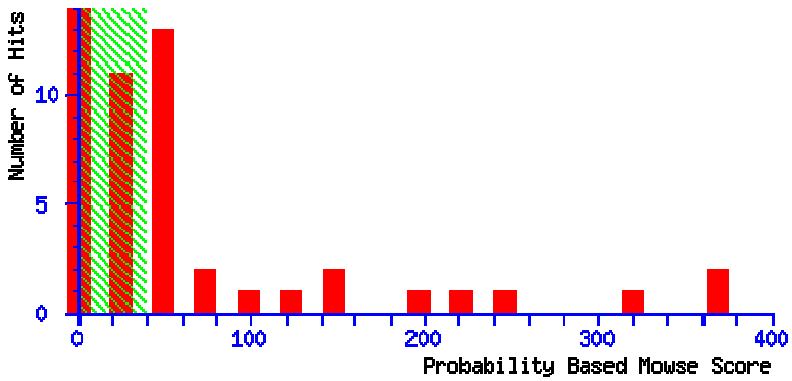
Výsledky analýzy (Mascot Search Results)

Tabulka 44: *Hlavička nalezených proteinů.*

```
Název úlohy     : Konvertovaný soubor z mzXML do MGF
Search title    : Conversion of 7mix2.mzXML to mascot generic

Název MS datového souboru,
MS data file    : 7mix2.mgf

Databáze vzorů,
Database        : IPI_HUMAN (69731 sequences;29254227 residues)

Datum a rok,
Timestamp       : 26 Feb 2008

Významné výsledky,
Significant hits:

IPI00025879  Tax_Id=9606 Gene_Symbol=MYH1 Myosin-1
IPI00001753  Tax_Id=9606 Gene_Symbol=MYH4 Myosin-4
IPI00007856  Tax_Id=9606 Gene_Symbol=MYH2 Myosin-2
IPI00298301  Tax_Id=9606 Gene_Symbol=MYH3 Myosin-3
IPI00302329  Tax_Id=9606 Gene_Symbol=MYH8 Myosin-8
IPI00218130  Tax_Id=9606 Gene_Symbol=PYGM Glycogen phosphorylase
IPI00302328  Tax_Id=9606 Gene_Symbol=MYH6 myosin heavy chain 6
IPI00555997  Tax_Id=9606 Gene_Symbol=MYH2 MYH2 protein
IPI00465436  Tax_Id=9606 Gene_Symbol=CAT  Catalase
...
IPI00514086  Tax_Id=9606 Gene_Symbol=TXNDC8 Uncharacterized protein
```

Probability Based Mowse Score

Iontové skóre je spočítáno ze vzorce $-10 \cdot$, kde P je hodnota pravděpodobnosti, že pozorovaná shoda (match) je náhodný jev. Individuální iontové skóre (ion skór) > 39 indikuje shodnost, nebo vysokou homologii s ($p<0.05$). *Proteinové skóre* je odvozeno z iontového skóre jako nepravděpodobnostní základ pro seřazení proteinových shod (hits).



Obrázek 27: *Sestavený histogram skóre.*

Souhrnná zpráva z vyhledávání peptidů (Peptide Summary Report)

1. IPI00025879

Tabulka 45: *Hlavička seznamu peptidů identifikovaného proteinu.*

Hmotnost, Ion-skóre, Počet souhlasných peptidů, **Mass:** 222976 **Score:** 368 **Peptides matched:** 49 Tax_Id=9606 Gene_Symbol=MYH1 Myosin-1

Tabulka 46: *Seznam peptidů pro identifikovaný protein.*

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
221	487.97	486.96	488.31	-1.34	1	(6)	3e+002	7	VKSR
222	488.00	486.99	488.31	-1.31	1	9	1.6e+002	2	VKSR
397	540.20	539.19	539.31	-0.12	0	16	9.9	1	VFFK
470	571.22	570.21	569.32	0.89	0	7	1e+002	2	LYFK
516	588.32	587.31	587.30	0.01	0	2	7.2e+002	1	NDAIR
669	651.15	650.14	649.31	0.83	0	11	40	3	CASLEK
786	709.23	708.22	708.36	-0.14	0	23	2.3	1	AFMNVK
...	...	...	...	...	...	...	...	...	...
2793	663.76	1988.26	1986.97	1.28	1	(16)	6.2	3	LEQQVDDLEGSLEQEKK
2973	1117.31	2232.61	2232.18	0.43	1	10	20	1	IEA…VDPK

Dále bude rozebrána položka 221 (query) a její podrobnější zobrazení označuje reziduum konkrétního peptidu „VKSR“ s možnými modifikacemi (vždy ale s jedinečným ID z databáze vzorových peptidů), které bylo nalezeno na deseti různých místech analyzované látky. Hodnota Ion skóre je uvedena v závorce, protože je stejný peptidový řetězec, který se vyskytuje jinde a má lepší skóre. Hodnota kvality (rank) je 7 (nejkvalitnější má 1), proto není zvýrazněn červenou barvou.

Skóre lepší než 21 zde indikuje podobnost (homologii), skóre lepší než 43 indikuje identitu. Číslo skenu je 428.

Query: 221

Tabulka 47: *Popis deseti modifikací jednoho peptidu pro položku 221.*

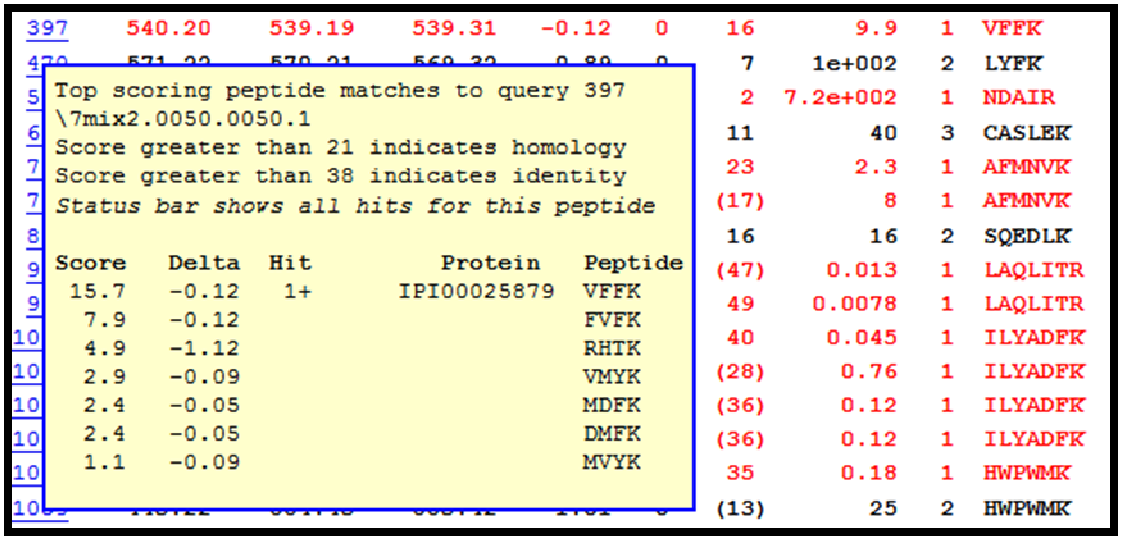
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
221	487.97	486.96	488.31	-1.34	1	(6)	3e+002	7	VKSR
						9	1.6e+002	2	VKSR
						16	9.9	1	VFFK
						7	1e+002	2	LYFK
						2	7.2e+002	1	NDAIR
						11	40	3	CASLEK
						23	2.3	1	AFMNVK
						(17)	8	1	AFMNVK
						16	16	2	SQEDLK
						(47)	0.013	1	LAQLITR
						49	0.0078	1	LAQLITR
						40	0.045	1	ILYADFK
						(28)	0.76	1	ILYADFK
						(36)	0.12	1	ILYADFK
						(36)	0.12	1	ILYADFK
						35	0.18	1	HWPWMK
						(13)	25	2	HWPWMK
						36	0.13	1	ANSEVAQWR
1399	531.75	1061.49	1059.51	1.98	0	(31)	0.38	1	ANSEVAQWR

Top scoring peptide matches to query 221
\7mix2.0428.0428.1
Score greater than 21 indicates homology
Score greater than 43 indicates identity
Status bar shows all hits for this peptide

	Score	Delta	Hit	Protein	Peptide
01	14.7	-1.30			NLSVG
02	7.6	0.65			KVALG
03	6.2	-1.31			GGLSR
04	6.2	-1.31			GLGSR
05	6.2	-1.31			GGISR
06	6.2	-1.31			GIGSR
07	5.8	-1.32			ARSR
08	5.8	-1.31			QVSR
09	5.8	-1.34	1+	IPI00025879	VKSR
10	5.8	-1.31			LNSR

Query 397

Tabulka 48: *Výpis podrobností pro položku 397 ze seznamu výsledků.*



397	540.20	539.19	539.31	-0.12	0	16	9.9	1	VFFK
470	571.22	570.21	569.32	0.89	0	7	1e+002	2	LYFK
5						2	7.2e+002	1	NDAIR
6						11	40	3	CASLEK
7						23	2.3	1	AFMNVK
7						(17)	8	1	AFMNVK
8						16	16	2	SQEDLK
9						(47)	0.013	1	LAQLITR
9						49	0.0078	1	LAQLITR
10						40	0.045	1	ILYADFK
10						(28)	0.76	1	ILYADFK
10						(36)	0.12	1	ILYADFK
10						(36)	0.12	1	ILYADFK
10						35	0.18	1	HWPWMK
10						(13)	25	2	HWPWMK

Top scoring peptide matches to query 397
\7mix2.0050.0050.1
Score greater than 21 indicates homology
Score greater than 38 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
15.7	-0.12	1+	IPI00025879	VFFK
7.9	-0.12			FVFK
4.9	-1.12			RHIK
2.9	-0.09			VMYK
2.4	-0.05			MDFK
2.4	-0.05			DMFK
1.1	-0.09			MVYK

Tabulka 49: *Výpis podrobností pro položku (Query) 397 z XML souboru.*

```
<spectrum_query spectrum="7mix2.0050.0050.1" start_scan="0050"
end_scan="0050" precursor_neutral_mass="539.1927" assumed_charge="1"
index="256" retention_time_sec="75.740000">
<search_result>
    <search_hit hit_rank="1" peptide="VFFK" peptide_prev_aa="K"
peptide_next_aa="A" protein="IPI:IPI00001753.1|SWISS-
PROT:Q9Y623|ENSEMBL:ENSP00000255381|REFSEQ:NP_060003|H-
INV:HIT000069701|VEGA:OTTHUMP00000160899;OTTHUMP00000183341"
num_tot_proteins="51" num_matched_ions="4" tot_num_ions="6"
calc_neutral_pep_mass="539.3107" massdiff="-0.1" num_tol_term="2"
num_missed_cleavages="0" is_rejected="0">
        <search_score name="ionscore" value="15.65"/>
        <search_score name="identityscore" value="38.62"/>
        <search_score name="star" value="0"/>
        <search_score name="homologyscore" value="21.92"/>
        <search_score name="expect" value="9.91"/>
    </search_hit>
</search_result>
</spectrum_query>
```

---

2. IPI00001753

Tabulka 50: *Identifikace druhého proteinu.*

Mass: 222874 **Score:** 358 **Peptides matched:** 53
Tax_Id=9606 Gene_Symbol=MYH4 Myosin-4

Tabulka 51: *Výpis nalezených peptidů druhého proteinu.*

Query	Observed	Mr (expt)	Mr (calc)	Delta	Miss	Score	Expect	Rank	Peptide
9	403.85	402.84	401.28	1.57	1	6	4.1e+002	1	RVK
10	404.01	403.00	401.28	1.73	1	(6)	4.2e+002	1	RVK
221	487.97	486.96	488.31	-1.34	1	(6)	3e+002	7	VKSR
222	488.00	486.99	488.31	-1.31	1	9	1.6e+002	2	VKSR
397	540.20	539.19	539.31	-0.12	0	16	9.9	1	VFFK
...	...	...	...	...	...	...	...	...	...
2434	560.63	1678.87	1680.77	-1.90	0	6	79	6	MEINDLA SNMETVSK
2457	565.71	1694.11	1695.95	-1.84	1	22	2.3	1	DLQLRLG EAEQLALK
2792	663.47	1987.39	1986.97	0.41	1	34	0.093	1	LEQQVDDL EGSLEQEKK
2793	663.76	1988.26	1986.97	1.28	1	(16)	6.2	3	LEQQVDDL EGSLEQEKK
2973	1117.31	2232.61	2232.18	0.43	1	10	20	1	IEAQNKPFDA KTSVFVVDPK

20. IPI00514086

Tabulka 52: *Identifikace posledního určeného proteinu.*

Mass: 12293 **Score:** 41 **Peptides matched:** 3
Tax_Id=9606 Gene_Symbol=TXNDC8 Uncharacterized protein TXNDC8

Tabulka 53: *Seznam nalezených peptidů k poslednímu proteinu.*

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
822	731.13	730.12	730.44	-0.32	0	41	0.062	1	MVQIIK
823	731.33	730.32	730.44	-0.12	0	(36)	0.22	1	MVQIIK
2957	1107.21	2212.41	2214.15	-1.75	1	8	35	6	SQKGSISRPT PTSELHPHQK

Peptidové shody, nepřiřazené žádnému proteinu

V originále uvedeném na výstupu analýzy (Peptide matches not assigned to protein hits): Neuvedené detaily znamenají absenci shody(no details means no match). Seznam je řazen vzhledem ke skóre a sestupně (List sorted by Decreasing Score).

Tabulka 54: *Neidentifikované fragmenty proteinů (peptidy).*

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1	400.52	399.51	399.18	0.33	0	10	1e+002	1	YAF
...	...	...	...	...	...	...	...	...	...
2669	933.55	1865.09	1866.92	1.84	0	0	2.6e+002	1	SGGYLW
5	*401.69*	*400.68*							
...	...	...	...	...	...	...	...	...	...

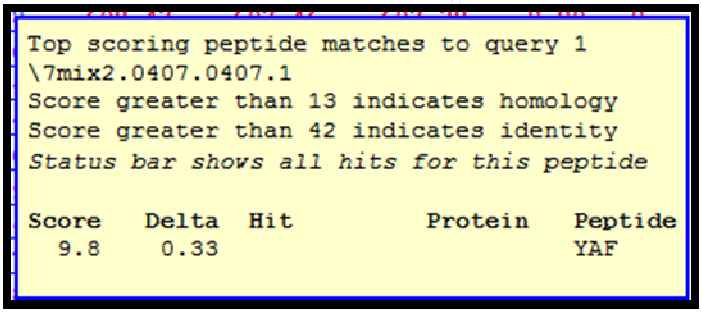
V následující tabulce je zobrazena položka 1 (index), jak je zapsaná v XML kódu souboru dat. Číslování souhlasí s prvním dotazem Query 1 v tabulce jen ze začátku.

Tabulka 55: *Výpis kódu ze souboru – způsob zápisu hitu Query 1 v XML jazyce.*

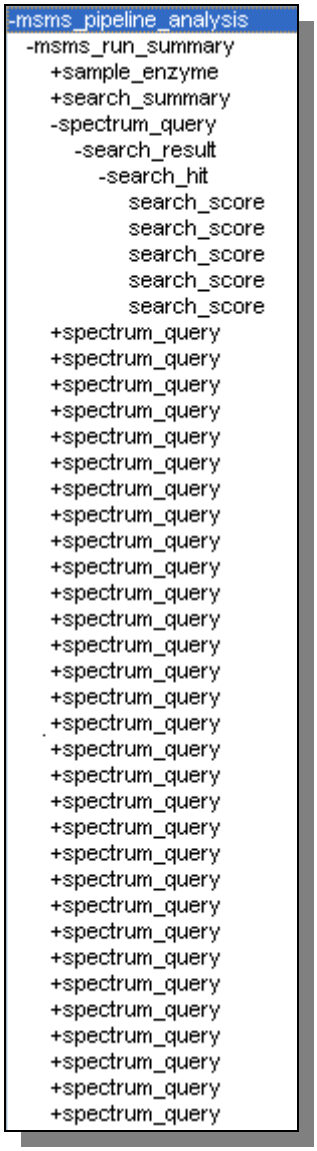
```
<spectrum_query spectrum="7mix2.0407.0407.1" start_scan="0407"
end_scan="0407" precursor_neutral_mass="399.5127" assumed_charge="1"
index="1" retention_time_sec="610.710000">
<search_result>
    <search_hit hit_rank="1" peptide="YAF" peptide_prev_aa="R"
peptide_next_aa="-" protein="IPI:IPI00300371.5|SWISS-PROT:Q15393-
1|TREMBL:A8K6V3|ENSEMBL:ENSP00000305790|REFSEQ:NP_036558|H-
INV:HIT000243874|VEGA:OTTHUMP00000081478;OTTHUMP00000174907"
num_tot_proteins="4" num_matched_ions="2" tot_num_ions="4"
calc_neutral_pep_mass="399.1794" massdiff="+0.3" num_tol_term="2"
num_missed_cleavages="0" is_rejected="0">
        <search_score name="ionscore" value="9.75"/>
        <search_score name="identityscore" value="42.77"/>
        <search_score name="star" value="0"/>
        <search_score name="homologyscore" value="13.61"/>
        <search_score name="expect" value="100.17"/>
    </search_hit>
</search_result>
</spectrum_query>
```

Pro neučenou položku je hranice pro skóre homologie 13 a identity 42. Start sken na hodnotě 0407 a koncový sken na pozici 0407.

Tabulka 56: *Submenu pro peptid položky Query 1.*



Tabulka 57: *Struktura dat položky Query 1*



Parametry analýzy (Search Parameters):

- Typ vyhledávání - Type of search: MS/MS Ion Search
- Enzym pro štěpení - Enzyme: Trypsin
- Typ hmotnosti - Mass values: Monoisotopic
- Hmotnost proteinu - Protein Mass: Unrestricted
- Tolerance hmotností pro peptidy - Peptide Mass Tolerance: ± 2 Da
- Tolerance hmotností pro fragmenty - Fragment Mass Tol.: ± 0.8 Da
- Maximální počet děr - Max Missed Cleavages: 1
- Typ nástroje - Instrument type: Default
- Název datového souboru - Data File Name: 7mix2.mgf
- Celkový počet výsledných položek - Number of queries: 3239

2.15.3 Výstup z X!Tandem

Vstupní parametry pro výpočet X!Tandem se nachází ke konci datového XML souboru. Zvolené parametry zastřešuje skupina (group) s názvem *vstupní parametry* (input parameters). K výpisu parametrů z obsáhlého datového souboru byly zvoleny tyto hodnoty:

Tabulka 58: *Výpis parametrů ze souboru „7mix2.tandem.xml“.*

```
<group label="input parameters" type="parameters">
<note type="input" label="list path, default parameters">
default_input.xml</note>
<note type="input" label="list path, taxonomy information">
taxonomy.xml</note>
<note type="input" label="output, maximum valid expectation value">
0.1</note>
<note type="input" label="output, one sequence copy">no</note>
<note type="input" label="output, path">7mix2.tandem.xml</note>
<note type="input" label="output, proteins">yes</note>
<note type="input" label="output, sequences">yes</note>
<note type="input" label="output, sort results by">protein</note>
<note type="input" label="output, xsl path">tandem-style.xsl</note>
<note type="input" label="protein, taxon">IPI_HUMAN</note>
<note type="input" label="refine, maximum valid expectation value">
0.1</note>
<note type="input" label="scoring, b ions">yes</note>
<note type="input" label="scoring, minimum ion count">4</note>
<note type="input" label="scoring, y ions">yes</note>
<note type="input" label="spectrum, fragment monoisotopic mass
error">
0.4</note>
<note type="input" label="spectrum, fragment monoisotopic mass error
units">
Da</note>
<note type="input" label="spectrum, path">7mix2.mzXML</note>
<note type="input" label="spectrum, total peaks">50</note>
</group>
```

Ukázka struktury formátu dat ze souboru <7mix2.tandem.xml> po převedení do Matlabu pomocí funkce *xml_read* vytvořený Jarkem Tuszynskim z SAIC. Tato externí funkce Matlabu načte data do proměnné typu struktura. Následně se použije další externí funkce od stejného autora. Má název *gen_object_display(“struktura”)* a slouží k zobrazení již načtené struktury.

Skupina (group): obsah skupiny je 153 buněk typu struktura [153x1 struct]

Tabulka 59: *Příklad hierarchie výsledků X!Tandem v datovém typu struktura po konverzi do Matlabu.*

<cell> 1 of 153 report
1. protein: [1x6 struct]

- 1.1. <cell> peptid
  - 1.1.1. note: [1x1 struct]
    - 1.1.1.1. CONTENT: [1x176 char]
    - 1.1.1.2. ATTRIBUTE: [1x1 struct]
      - 1.1.1.2.1. label: 'description'
  - 1.1.2. file: [1x1 struct]
    - 1.1.2.1. CONTENT: [0x0 double]
    - 1.1.2.2. ATTRIBUTE: [1x1 struct]
      - 1.1.2.2.1. URL: 'IPI_Human.fasta'
      - 1.1.2.2.2. type: 'peptide'
  - 1.1.3. peptide: [1x1 struct]
    - 1.1.3.1. CONTENT: [1x2170 char]
    - 1.1.3.2. domain: [1x1 struct]
      - 1.1.3.2.1. CONTENT: [0x0 double]
      - 1.1.3.2.2. ATTRIBUTE: [1x1 struct]
        - 1.1.3.2.2.1. b_ions: [4]
        - 1.1.3.2.2.2. b_score: [12.3]
        - 1.1.3.2.2.3. delta: [0.655]
        - 1.1.3.2.2.4. xEnd: [91]
        - 1.1.3.2.2.5. expect: [0.0035]
        - 1.1.3.2.2.6. hyperscore: [30.8]
        - 1.1.3.2.2.7. id: '732.1.1'
        - 1.1.3.2.2.8. mh: [913.398]
        - 1.1.3.2.2.9. missed_cleavages: [6]
        - 1.1.3.2.2.10. nextscore: [18.4]
        - 1.1.3.2.2.11. post: 'AMMT'
        - 1.1.3.2.2.12. pre: 'NPPK'
        - 1.1.3.2.2.13. seq: 'YDKIEDM'
        - 1.1.3.2.2.14. start: [85]
        - 1.1.3.2.2.15. y_ions: [6]
        - 1.1.3.2.2.16. y_score: [12.9]
    - 1.1.3.3. ATTRIBUTE: [1x1 struct]
      - 1.1.3.3.1. xEnd: [1939]
      - 1.1.3.3.2. start: [1]
  - 1.1.4. ATTRIBUTE: [1x1 struct]
    - 1.1.4.1. expect: [-166.7]
    - 1.1.4.2. id: [732.1]

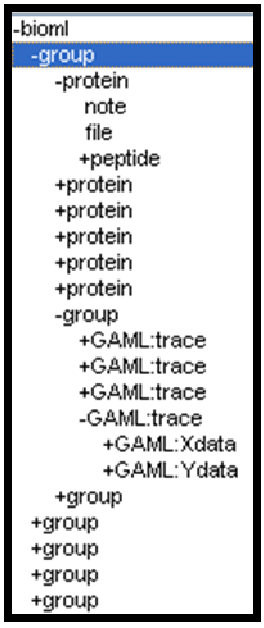
- 1.1.4.3. label: [1x141 char]
- 1.1.4.4. sumI: [9]
- 1.1.4.5. uid: [6986]
- 1.2. <cell> peptid
- 1.3. <cell> peptid
- 1.4. <cell> peptid
- 1.5. <cell> peptid
- 1.6. <cell> peptid

- 2. group: [2x1 struct]
  - 2.1. <cell>
    - 2.1.1. GAML_COLON_trace: [4x1 struct]
      - 2.1.1.1. <cell>
        - 2.1.1.1.1. GAML_COLON_attribute: [1x2 struct]
          - 2.1.1.1.1.1. <cell>
            - 2.1.1.1.1.1.1. CONTENT: [6.1253]
            - 2.1.1.1.1.1.2. ATTRIBUTE: [1x1 struct]
              - 2.1.1.1.1.1.2.1. type: 'a0'
          - 2.1.1.1.1.2. <cell>
            - 2.1.1.1.1.2.1. CONTENT: [-0.27852]
            - 2.1.1.1.1.2.2. ATTRIBUTE: [1x1 struct]
              - 2.1.1.1.1.2.2.1. type: 'a1'
        - 2.1.1.1.2. GAML_COLON_Xdata: [1x1 struct]
          - 2.1.1.1.2.1.1. GAML_COLON_values: [1x1 struct]
            - 2.1.1.1.2.1.1.1. CONTENT: [1x88 char]
            - 2.1.1.1.2.1.1.2. ATTRIBUTE: [1x1 struct]
          - 2.1.1.1.2.1.2. ATTRIBUTE: [1x1 struct]
            - 2.1.1.1.2.1.2.1. label: '732.hyper'
            - 2.1.1.1.2.1.2.2. units: 'score'
        - 2.1.1.1.3. GAML_COLON_Ydata: [1x1 struct]
          - 2.1.1.1.3.1. GAML_COLON_values: [1x1 struct]
            - 2.1.1.1.3.1.1. CONTENT: [1x117 char]
            - 2.1.1.1.3.1.2. ATTRIBUTE: [1x1 struct]
          - 2.1.1.1.3.2. ATTRIBUTE: [1x1 struct]
            - 2.1.1.1.3.2.1. label: '732.hyper'
            - 2.1.1.1.3.2.2. units: 'counts'
        - 2.1.1.1.4. ATTRIBUTE: [1x1 struct]

2.1.1.1.4.1.	label: '732.hyper'
2.1.1.1.4.2.	type: 'hyperscore expectation function'
2.1.2.	

Další toolbox, s názvem *XMLTree ver.1* od *Guillaume Flandina* dovoluje zobrazení struktury ve formě stromu. Pro zobrazení všech položek však není prostor, proto je ve stromové struktuře zobrazeno jen několik skupin.

Obrázek 28: *View X!Tandem pro 150 skupin (group).*



X!Tandem umožňuje on-line prohlížení dat změřených na www stránce v grafické úpravě. Výsledkem X!Tandem analýzy je 20 identifikovaných proteinů a 35 dalších proteinů, označených jako podobné:

Výstupní bilance X!Tandem:

20 proteinů + 35 podobností k proteinům = 55 celkem nalezených proteinů (20 proteins + 35 homologs = 55 total proteins)

Tabulka 60: *Seznam identifikovaných proteinů.*

RANK	LOG(E)$^{+}$	LOG(I)	%/%	#	TOTAL	MR	ACCESSION
1	-166.7	9.00	14/21	32	50	223.0	IPI:IPI00025879.1
2	-138.3	8.95	3.2/5	6	8	222.9	(H) IPI:IPI00001753.1
3	-137.1	9.37	21/29	30	55	97.0	IPI:IPI00218130.3
4	-129.8	8.96	0.5/1	1	2	222.9	(H) IPI:IPI00007856.1
...	...	...	...	...	...	...	...
19	-1.1	7.00	2.2/4	1	1	70.5	IPI:IPI00025489.1
20	-1.1	7.27	1.0/2	1	1	107.0	IPI:IPI00027744.1

Významy jednotlivých sloupců jsou následující:

- Rank – kvalitativní hodnocení podle zvoleného kritéria řazení v seznamu.
- Log (e) – hodnota expectation v logaritmu o základu -10.
- Log (I) – hodnota intenzity v logaritmu o základu -10.
- %/% - překrytí proteinu.
- # - počet nalezených peptidů v proteinu.
- Total – celkový počet nalezených peptidů.
- Mr – molekulová hmotnost proteinu v kilodaltonech.
- Accession – přístupový identifikační ID kód v databázi vzorových proteinů.

Nyní je uvedena první položka XML souboru jako výstupu X!Tandem.

Tabulka 61: *XML verze MYOSIN-1*

<protein expect="-166.7" id="732.1" uid="6986" label="IPI:**IPI00025879.1**|SWISS-PROT:P12882|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954|H-INV:HIT000069703|VEGA:OTTHUMP00000160893;OTTHUMP00000183342..." sumI="9.00" >

Stejná položka v provedení on-line webové stánky:

Tabulka 62: *Hodnota expect s identifikačním kódem v obdobných databázích.*

log(e) = -166.7 **IPI:IPI00025879.1|**
SWISS-PROT:P12882|ENSEMBL:ENSP00000226207|REFSEQ:NP_005954|
H-INV:HIT000069703|VEGA:OTTHUMP00000160893

Tabulka 63: *Zvýrazněné identifikované peptidy v nalezeném proteinu MYOSIN-1*

```
  1 MSSDSEMAIFGEAAPFLRKSERERIEAQNKPFDAKTSVFVVDPKESFVKATVQSREGGKV  60
 61 TAKTEAGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLKERYAAWMIYTYSG 120
121 LFCVTVNPYKWLPVYNAEVVTAYRGKKRQEAPPHIFSISDNAYQFMLTDRENQSILITGE 180
181 SGAGKTVNTKRVIQYFATIAVTGEKKKEEVTSGKMQGTLEDQIISANPLLEAFGNAKTVR 240
241 NDNSSRFGKFIRIHFGTTGKLASADIETYLLEKSRVTFQLKAERSYHIFYQIMSNKKPDL 300
301 IEMLLITTNPYDYAFVSQGEITVPSIDDQEELMATDSAIEILGFTSDERVSIYKLTGAVM 360
361 HYGNMKFKQKQREEQAEPDGTEVADKAAYLQNLNSADLLKALCYPRVKVGNEYVTKGQTV 420
421 QQVYNAVGALAKAVYDKMFLWMVTRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLC 480
481 INFTNEKLQQFFNHHMFVLEQEEYKKEGIEWTFIDFGMDLAACIELIEKPMGIFSILEEE 540
541 CMFPKATDTSFKNKLYEQHLGKSNNFQKPKPAKGKPEAHFSLIHYAGTVDYNIAGWLDKN 600
601 KDPLNETVVGLYQKSAMKTLALLFVGATGAEAEAGGGKKGGKKKGSSFQTVSALFRENLN 660
661 KLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVLHQLRCNGVLEGIRICRKGFPSRILYA 720
721 DFKQRYKVLNASAIPEGQFIDSKKASEKLLGSIDIDHTQYKFGHTKVFFKAGLLGLLEEM 780
781 RDEKLAQLITRTQAMCRGFLARVEYQKMVERRESIFCIQYNVRAFMNVKHWPWMKLYFKI 840
841 KPLLKSAETEKEMANMKEEFEKTKEELAKTEAKRKELEEKMVTLMQEKNDLQLQVQAEAD 900
901 SLADAEERCDQLIKTKIQLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDD 960
```

```
 961 LELTLAKVEKEKHATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEEDK 1020
1021 VNTLTKAKIKLEQQVDDLEGSLEQEKKIRMDLERAKRKLEGDLKLAQESAMDIENDKQQL 1080
1081 DEKLKKKEFEMSGLQSKIEDEQALGMQLQKKIKELQARIEELEEEIEAERASRAKAEKQR 1140
1141 SDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDLEEATLQHEATAATLRKKH 1200
1201 ADSVAELGEQIDNLQRVKQKLEKEKSEMKMEIDDLASNMETVSKAKGNLEKMCRALEDQL 1260
1261 SEIKTKEEEQQRLINDLTAQRARLQTESGEYSRQLDEKDTLVSQLSRGKQAFTQQIEELK 1320
1321 RQLEEEIKAKSALAHALQSSRHDCDLLREQYEEEQEAKAELQRAMSKANSEVAQWRTKYE 1380
1381 TDAIQRTEELEEAKKKLAQRLQDAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDVERTN 1440
1441 AACAALDKKQRNFDKILAEWKQKCEETHAELEASQKESRSLSTELFKIKNAYEESLDQLE 1500
1501 TLKRENKNLQQEISDLTEQIAEGGKRIHELEKIKKQVEQEKSELQAALEEAEASLEHEEG 1560
1561 KILRIQLELNQVKSEVDRKIAEKDEEIDQMKRNHIRIVESMQSTLDAEIRSRNDAIRLKK 1620
1621 KMEGDLNEMEIQLNHANRMAAEALRNYRNTQAILKDTQLHLDDALRSQEDLKEQLAMVER 1680
1681 RANLLQAEIEELRATLEQTERSRKIAEQELLDASERVQLLHTQNTSLINTKKKLETDISQ 1740
1741 IQGEMEDIIQEARNAEEKAKKAITDAAMMAEELKKEQDTSAHLERMKKNLEQTVKDLQHR 1800
1801 LDEAEQLALKGGKKQIQKLEARVRELEGEVESEQKRNVEAVKGLRKHERKVKELTYQTEE 1860
1861 DRKNILRLQDLVDKLQAKVKSYKRQAEEAEEQSNVNLSKFRRIQHELEEAEERADIAESQ 1920
1921 VNKLRVKSREVHTKIISEE                                         1939
```

Identifikované peptidy (Identified Peptides) proteinu MYOSIN 1 a jejich podrobný popis: **50 položek.**

Tabulka 64: *Podrobný popis nalezených peptidových sekvencí proteinu MYOSIN1.*

spektrum spectrum	log(e)	log(I)	m+h	delta	z	sekvence semence	
732.1	-2.5	6.80	913.398	0.655	2	nppk[85]	YDKIEDM [91]ammt (0)
1430.1	-3.8	7.54	1179.616	0.517	2	mamm[95]	THLHEPAVLY [104]nlke (30)
...	...	...	...	...	...	...	...
441.1	-1.8	5.80	1831.878	0.727	3	ksyk[1884]	RQAEEAEEQS NVNLSK [1899]frri (0)
599.1	-8.4	6.98	1675.777	0.375	2	sykr[1885]	QAEEAEEQSN VNLSK [1899]frri (55)
720.1	-3.2	6.94	1382.655	0.358	2	kfrr[1903]	IQHELEEAEE R [1913]adia (104)

IPI protein report:

Tabulka 65: *Výpis podrobností reportu výstupu po on-line analýze X!Tandem*

```
ID   IPI00025879.1 IPI; PRT; 1939 AA.
AC   IPI00025879; IPI00103916;
DT   01-OCT-2001 (IPI Human rel. 2.00, Created)
DT   01-OCT-2001 (IPI Human rel. 2.00, Last sequence update)
DE   MYOSIN-1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
CC   -!- GENE_LOCATION: Chr. 17:10336354-10362584:-1.
DR   UniProtKB/Swiss-Prot; P12882; MYH1_HUMAN; M.
```

```
DR   Vega; OTTHUMP00000160893; OTTHUMG00000130362; -.
DR   Vega; OTTHUMP00000183342; OTTHUMG00000142740; -.
DR   REFSEQ_REVIEWED; NP_005954; GI:115527082; -.
DR   ENSEMBL; ENSP00000226207; ENSG00000109061; -.
DR   H-InvDB; HIT000069703; HIX0039062; -.
DR   UniParc; UPI000012FB6B; -; -.
DR   HGNC; 7567; MYH1; -.
DR   Entrez Gene; 4619; MYH1; -.
DR   UniGene; Hs.440895; -; -.
DR   CCDS; CCDS11155.1; -; -.
DR   trome; HTR008531; -; PRT.
DR   CleanEx; HS_MYH1; -; -.
DR   InterPro; IPR000048; IQ_CaM_bd_region.
DR   InterPro; IPR015650; Myosin_hc.
DR   InterPro; IPR001609; Myosin_head.
DR   InterPro; IPR004009; Myosin_N.
DR   InterPro; IPR002928; Myosin_tail.
DR   Pfam; PF00612; IQ; 1.
DR   Pfam; PF00063; Myosin_head; 1.
DR   Pfam; PF02736; Myosin_N; 1.
DR   Pfam; PF01576; Myosin_tail_1; 1.
DR   PRINTS; PR00193; MYOSINHEAVY.
DR   ProDom; PD000355; Myosin_head; 1.
DR   SMART; SM00015; IQ; 1.
DR   SMART; SM00242; MYSc; 1.
DR   PROSITE; PS50096; IQ; 1.
DR   PANTHER; PTHR13140:SF22; Myosin_hc; 1.
SQ   SEQUENCE   1939 AA;  223115 MW;  39ADB26AB79DFA53 CRC64;
```

Tabulka 66: *Export sekvence kompletního proteinu z databáze IPI human.*

```
     MSSDSEMAIF GEAAPFLRKS ERERIEAQNK PFDAKTSVFV VDPKESFVKA TVQSREGGKV
    TAKTEAGATV TVKDDQVFPM NPPKYDKIED MAMMTHLHEP AVLYNLKERY AAWMIYTYSG
    LFCVTVNPYK WLPVYNAEVV TAYRGKKRQE APPHIFSISD NAYQFMLTDR ENQSILITGE
    SGAGKTVNTK RVIQYFATIA VTGEKKKEEV TSGKMQGTLE DQIISANPLL EAFGNAKTVR
    NDNSSRFGKF IRIHFGTTGK LASADIETYL LEKSRVTFQL KAERSYHIFY QIMSNKKPDL
    IEMLLITTNP YDYAFVSQGE ITVPSIDDQE ELMATDSAIE ILGFTSDERV SIYKLTGAVM
    HYGNMKFKQK QREEQAEPDG TEVADKAAYL QNLNSADLLK ALCYPRVKVG NEYVTKGQTV
    QQVYNAVGAL AKAVYDKMFL WMVTRINQQL DTKQPRQYFI GVLDIAGFEI FDFNSLEQLC
    INFTNEKLQQ FFNHHMFVLE QEEYKKEGIE WTFIDFGMDL AACIELIEKP MGIFSILEEE
    CMFPKATDTS FKNKLYEQHL GKSNNFQKPK PAKGKPEAHF SLIHYAGTVD YNIAGWLDKN
    KDPLNETVVG LYQKSAMKTL ALLFVGATGA EAEAGGGKKG GKKKGSSFQT VSALFRENLN
    KLMTNLRSTH PHFVRCIIPN ETKTPGAMEH ELVLHQLRCN GVLEGIRICR KGFPSRILYA
    DFKQRYKVLN ASAIPEGQFI DSKKASEKLL GSIDIDHTQY KFGHTKVFFK AGLLGLLEEM
    RDEKLAQLIT RTQAMCRGFL ARVEYQKMVE RRESIFCIQY NVRAFMNVKH WPWMKLYFKI
    KPLLKSAETE KEMANMKEEF EKTKEELAKT EAKRKELEEK MVTLMQEKND LQLQVQAEAD
    SLADAEERCD QLIKTKIQLE AKIKEVTERA EDEEEINAEL TAKKRKLEDE CSELKKDIDD
    LELTLAKVEK EKHATENKVK NLTEEMAGLD ETIAKLTKEK KALQEAHQQT LDDLQAEEDK
```

```
     VNTLTKAKIK LEQQVDDLEG SLEQEKKIRM DLERAKRKLE GDLKLAQESA MDIENDKQQL
     DEKLKKKEFE MSGLQSKIED EQALGMQLQK KIKELQARIE ELEEEIEAER ASRAKAEKQR
     SDLSRELEEI SERLEEAGGA TSAQIEMNKK REAEFQKMRR DLEEATLQHE ATAATLRKKH
     ADSVAELGEQ IDNLQRVKQK LEKEKSEMKM EIDDLASNME TVSKAKGNLE KMCRALEDQL
     SEIKTKEEEQ QRLINDLTAQ RARLQTESGE YSRQLDEKDT LVSQLSRGKQ AFTQQIEELK
     RQLEEEIKAK SALAHALQSS RHDCDLLREQ YEEEQEAKAE LQRAMSKANS EVAQWRTKYE
     TDAIQRTEEL EEAKKKLAQR LQDAEEHVEA VNAKCASLEK TKQRLQNEVE DLMIDVERTN
     AACAALDKKQ RNFDKILAEW KQKCEETHAE LEASQKESRS LSTELFKIKN AYEESLDQLE
     TLKRENKNLQ QEISDLTEQI AEGGKRIHEL EKIKKQVEQE KSELQAALEE AEASLEHEEG
     KILRIQLELN QVKSEVDRKI AEKDEEIDQM KRNHIRIVES MQSTLDAEIR SRNDAIRLKK
     KMEGDLNEME IQLNHANRMA AEALRNYRNT QAILKDTQLH LDDALRSQED LKEQLAMVER
     RANLLQAEIE ELRATLEQTE RSRKIAEQEL LDASERVQLL HTQNTSLINT KKKLETDISQ
     IQGEMEDIIQ EARNAEEKAK KAITDAAMMA EELKKEQDTS AHLERMKKNL EQTVKDLQHR
     LDEAEQLALK GGKKQIQKLE ARVRELEGEV ESEQKRNVEA VKGLRKHERK VKELTYQTEE
     DRKNILRLQD LVDKLQAKVK SYKRQAEEAE EQSNVNLSKF RRIQHELEEA EERADIAESQ
     VNKLRVKSRE VHTKIISEE
//
```

Vzhledem k časové a výpočetní náročnosti zpracování velkých datových souborů se od skenu 2000 do konce 7161 pro testovací účely naměřená data ořezala. Posledních 5161 skenů je tedy na rozdíl od plné verze v testovací odstraněno. Plná verze naměřených raw dat je označena „7mix2_full“.

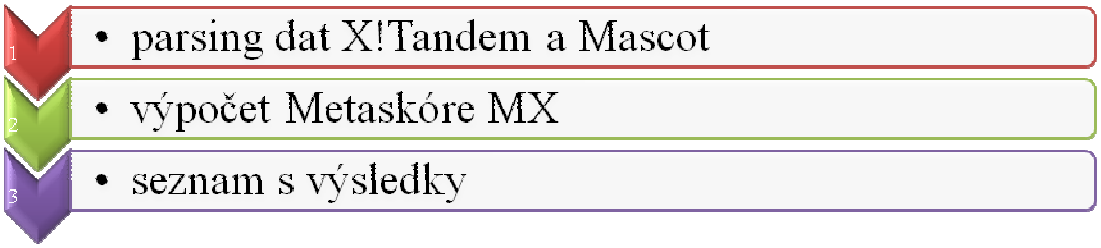
3. PRAKTICKÁ REALIZACE

3.1 Komunikace s programem v prostředí MATLAB

Program je realizovaný jako funkce z běžícího MATLABu. Vyvolání nápovědy pro přesnější parametry se spouští klíčovým slovem *help*. Podmínkou pro správné spuštění je korektní cesta s výsledkům z Mascot a X!Tandem analýzi ve formátu XML.

3.2 Algoritmus metaskóre

Skládá se ze tří praktických kroků. První je výtah dat z výstupních souborů X!Tandemu a Mascotu. Druhý je vlastní výpočet v Matlabu a třetí je prezentace výsledků, realizována také v prostředí Matlab.

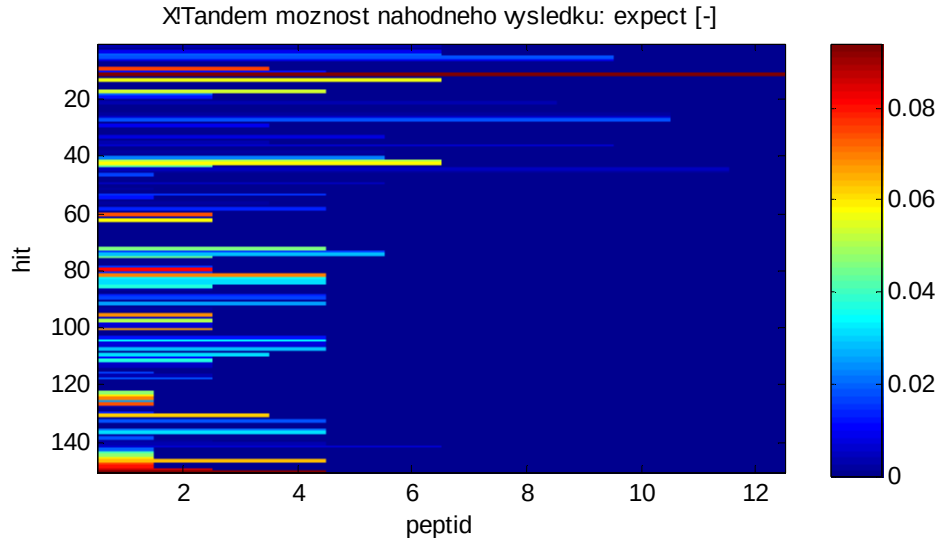


Obrázek 29: *Pracovní postup (workflow)*

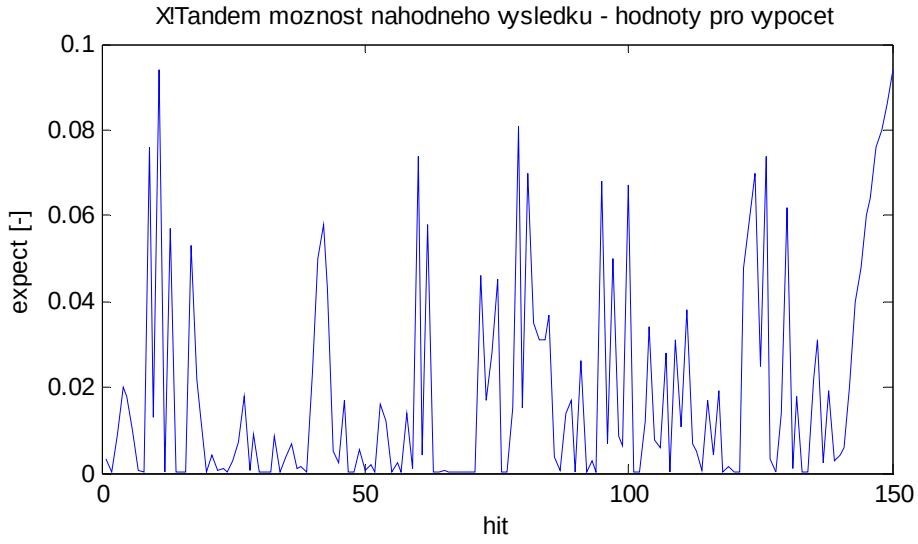
3.2.1 Parsování datových souborů

Před parsováním už musí být naměřená mzXML data poslána do Mascot a X!Tandem analýzi. Oba výstupy už musí také být připraveny ve formátu XML. Využívá se zde datového typu struktura, do které se vstupní textového formátu v jazyce XML převedou.

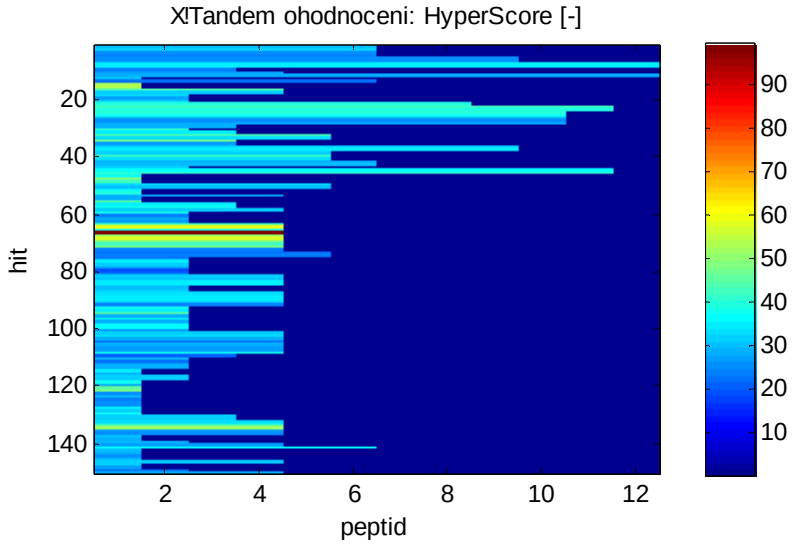
U X!Tandemu je systém uložených dat odlišný, než Mascotu. V X!Tandemu se je přiřazeno pro každý protein několik skupin. Všechny položky skupin všech identifikovaných proteinů jsou ve stromové struktuře uloženy do ploché databáze vedle sebe. Každá skupina obsahuje hitparádu proteinů, které odpovídají nalezeným fragmentům peptidů. Nejvhodnější peptid se označí za vedoucího a pojmenuje se po něm celá skupina. V Mascotu je situace jiná. Ve starších verzích byl pro konverzi z MGF do XML potřeba nástroj TPP. Po jeho použití byla data organizována do ploché databáze na všech úrovních, strom celé struktury byl pro analýzu těžce čitelný. Od Mascot verze 2.2 je k dispozici skript napsaný v jazyce Perl. Ten umožňuje přehledný export dat do XML souboru. Struktura obsahuje velice přehlednou a logickou hierarchii. Jako výchozí informace jsou parametry a identifikační údaje měření. Potom následuje přehledný seznam nalezených proteinů. Každý identifikovaný protein obsahuje jednu skupinu nalezených ohodnocených peptidů.



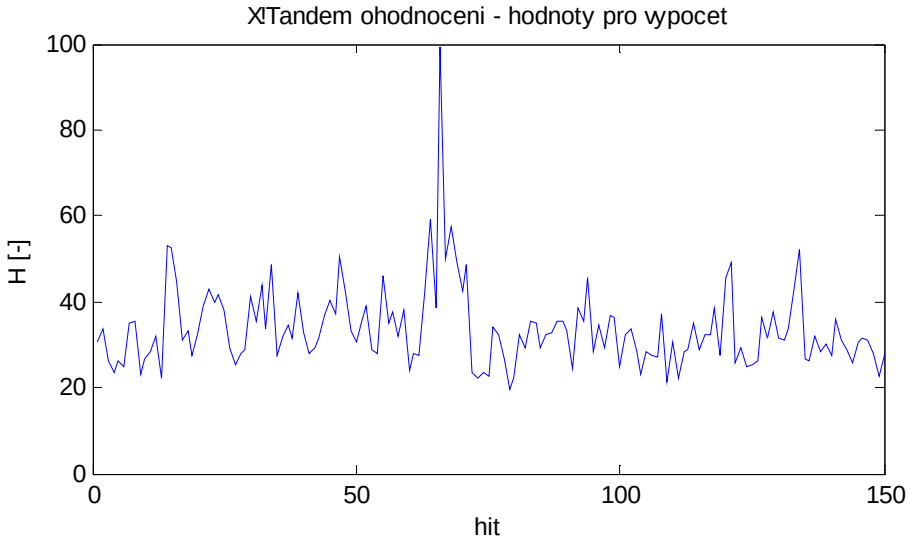
Obrázek 30: *X!Tandem - Přehled naměřených dat (hit groups) pro expect v doménách.*



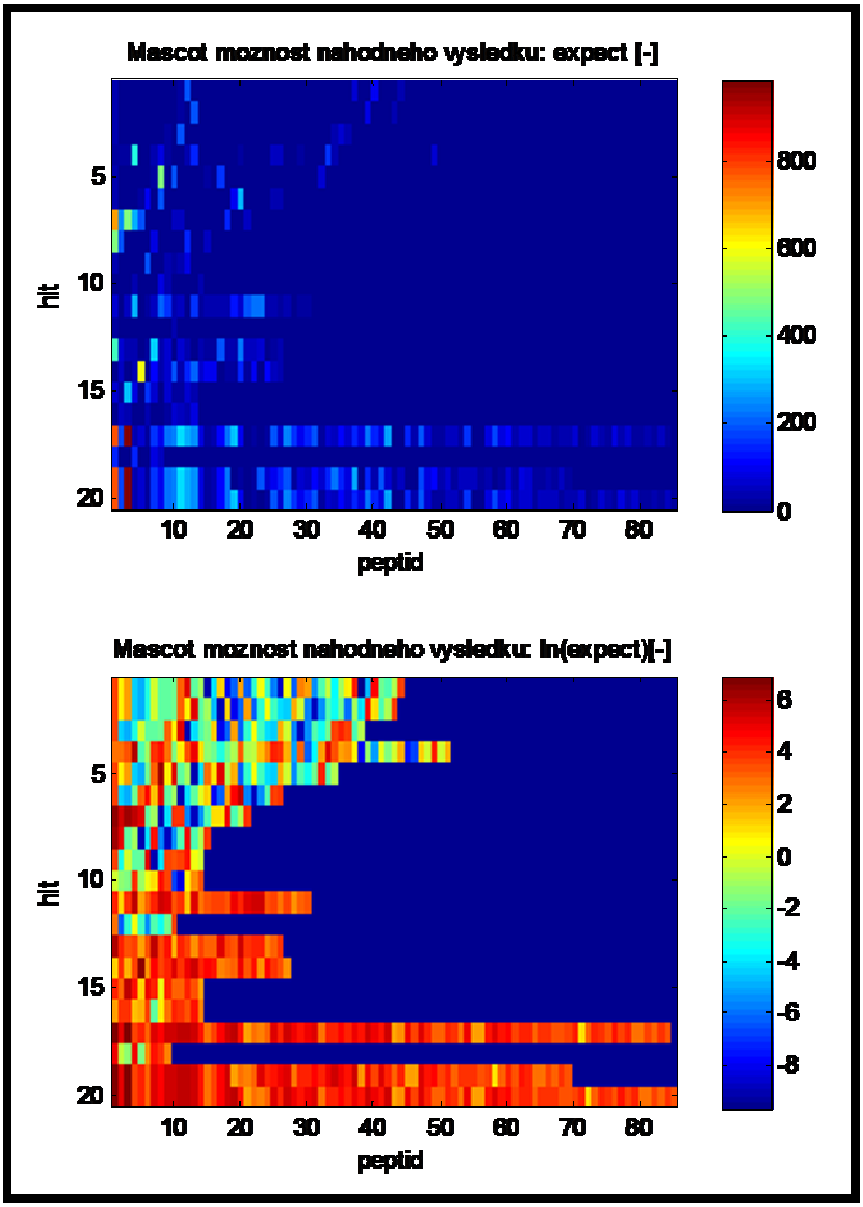
Obrázek 31: *X!Tandem - Naměřená data hodnot expectation upravená pro výpočet vybráním nejlepšího kandidáta z každé skupiny proteinů (peptidů).*



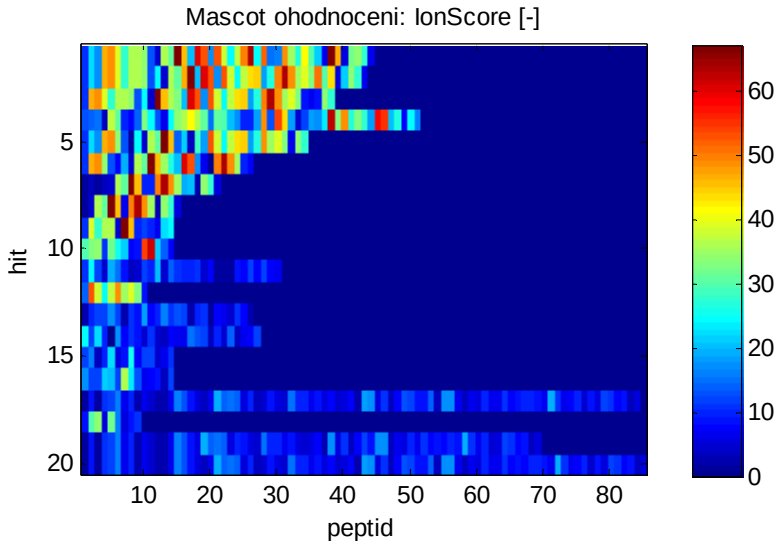
Obrázek 32: *X!Tandem - Přehled naměřených dat (hit groups) hyperskóre v doménách.*



Obrázek 33: *X!Tandem - Naměřená data hodnot expectation upravená pro výpočet vybráním nejlepšího kandidáta z každé skupiny proteinů (peptidů).*



Obrázek 34: *Výstup hodnot expectation pro nalezené proteiny v Mascotu (nahoře). Optimalizace stejných dat přirozeným logaritmem pro snížení dynamiky a zvýšení přehlednost (dole).*



Obrázek 35: *Výstup hodnot Ion-score z Mascot analýzi.*

3.2.2 Realizace metaskóre

Příklad výsledků pro první protein: **IPI:25879.1**, **Myosin-1** z testovací databáze 7mix2 a shodně nalezený oběma metodami na prvním místě.

Tabulka 67: *Příklad mezi-výsledků pro MX vybraného proteinu A.*

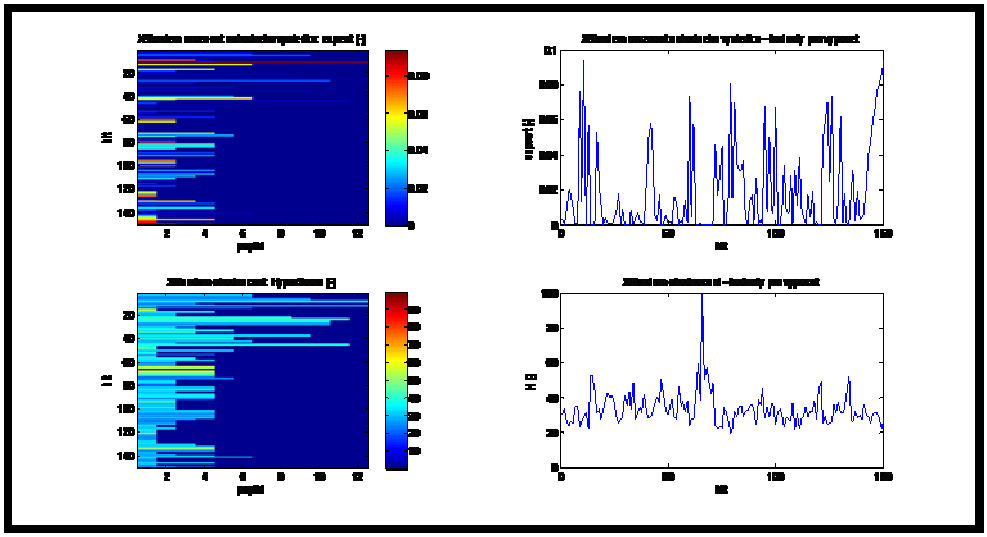
Protein A	X!Tandem			Mascot		
index	H	E_T	$H*E_T$	M	E_M	$M*E_T$
1	30,8	0,00350	0,10780	11,29	42,000	474,18
2	33,6	0,00016	0,00538	22,97	2,400	55,13
3	26,3	0,00900	0,23670	17,02	9,100	154,88
4	23,6	0,02000	0,47200	46,96	0,013	0,61
5	26,1	0,01800	0,46980	49,03	0,008	0,40
...	...	...	...	...	...	...
49	33,4	0,0057	0,19038	-	-	-
50	30,8	0,00069	0,021252	-	-	-

Tabulka 68: *Tabulka součtu příspěvků a výsledného metaskóre pro protein A.*

Protein: ID	X!Tandem: suma ($H*E_T$)	Mascot: suma ($M*E_T$)	Metaskóre: MX
A	18,75	3968,65	1993,70

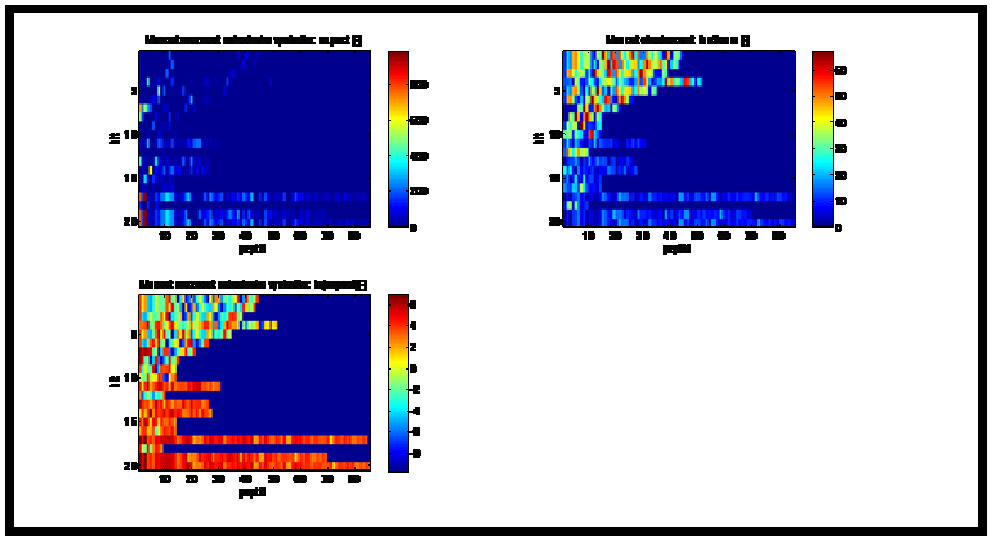
3.2.3 Reprezentace výstupu

Výstup programu je formou tří oken s grafy a jedné tabulky. První okno udává přehled o rozložení hodnocení vstupních dat X!Tandemu.



Obrázek 36: *Přehled vstupních hodnot z X!Tandemu.*

Druhé okno sdružuje vstupní hodnoty Mascotu. Grafy jsou popsané výše. Levá polovina patří expect hodnocení, pravá skóre.



Obrázek 37: *Přehled vstupních hodnot z Mascotu.*

Výstupní tabulka se skládá ze srovnání pořadí hodnocení X!Tandem, Mascot a Metaskóre. IPI kódu identifikovaných proteinů, a přiřazených ohodnocení v každé metodě. Pořadí výsledků jsou vždy vztaženy k Metaskóru (čím menší, tím věrohodnější identifikace).

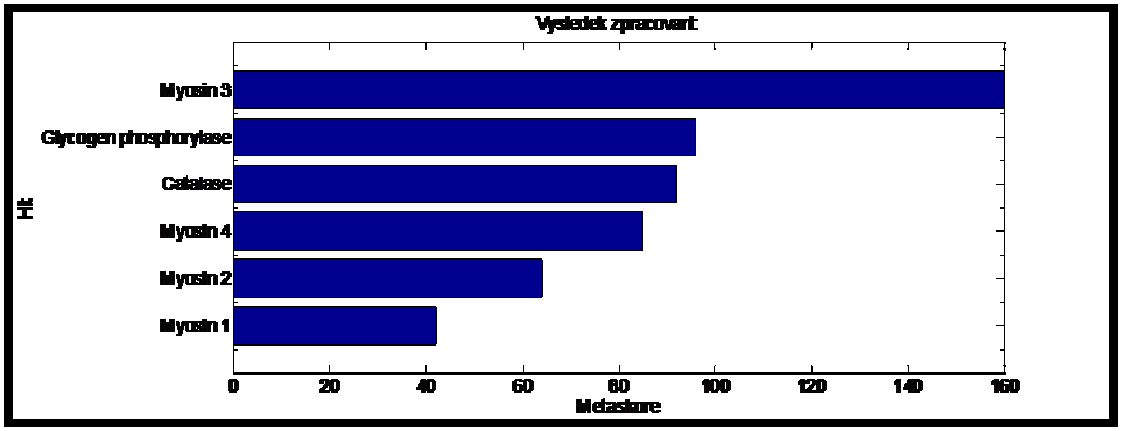
Tabulka 69: *Pořadí sjednocených dat testovacího souboru „7mix2“ od nejlepšího po nejhorší u všech třech metod.*

Metaskóre	X!Tandem	Mascot
1	1	1
2	2	2
3	4	3
4	6	10
5	3	4
6	5	6

Pro přehlednost je v dokumentaci tabulka výstupu skóre uvedena zvlášť. Také jsou zde doplněny názvy odpovídající ukazateli ID kódu v IPI databázi.

Tabulka 70: *Výsledek nalezených a ohodnocených proteinů numericky.*

Řazení	ID v IPI	Název	X!Tandem	Mascot	MX
1	25879	Myosin 1	0,01322	723	42
2	1753	Myosin-4	0,00588	697	64
3	7856	Myosin-2	0,00885	612	85
4	465436	Catalase	0,03084	168	92
5	218130	Glycogen phosphorylase	0,01895	539	96
6	298301	Myosin-3	0,01150	438	160



Obrázek 38: *Výsledek nalezených a ohodnocených proteinů graficky.*

3.3 Vyhodnocení metody

Metoda ohodnotila hlavní nalezené proteiny. Ze začátku je pořadí skórovaných shodné jak Mascot, X!Tandem i Metaskóre. Po cca. prvních třech pozicích se rozdílné přístupy začaly mírně rozcházet. Rozsah výstupů metaskóre je závislý na vstupních datech, u kterých je možná jakákoliv výstupní hodnota.

Výhodou aplikace metaskóre je zjednodušení mnoha dílčích výsledků do jednoho určujícího parametru. Sjednocením výsledků obou výchozích metod (Mascot, X!Tandem) také dochází k průniku stejné množiny dat. Tento průnik se stává věrohodným výsledkem, protože se na identifikaci podílely dva odlišné přístupy.

Nevýhodou je časová a výpočetní náročnost díky velkým datovým objemům a také zvoleným programovacím prostředím MATLAB.

4. ZÁVĚR

V první části práce je zahrnut teoretický úvod do genomiky a proteomiky. V další je tento teoretický základ použit pro identifikaci vzorku neznámé aminokyseliny. Jsou nastíněné dvě základní možnosti identifikace. Zaměření se provedlo na analýzu vzorků s pomocí použití databáze již části sekvenovaného proteomu. Fyzický rozbor vzorku byl proveden pomocí tandemové hmotnostní spektrometrie (MS/MS). Získaná MS/MS data jsou vstupem pro softwarové nástroje pro identifikací a zhodnocení kvality identifikace (skóre). Z možných nástrojů pro identifikaci byly vybrány dva, nekomerční X!Tandem a komerčně nabízený Mascot. U každé z těchto vybraných metod bylo popsáno jejich vstupní prostředí, schéma skórování a struktura datového formátu výstupních dat. Výstupy z obou metod jsou rozsáhlé. Také jsou u Mascot a X!Tandem vzájemně odlišné peptidové listy s mnoha nejednoznačnými položkami.

V praktické části byl tento problém byl řešen navržením systému pro zpracování výstupů a nahrazení rozdílného hodnocení jediným určujícím parametrem. Výsledkem je sjednocený seznam identifikovaných proteinů, s jedním určujícím parametrem pro každý protein. Tento parametr se nazývá Metaskóre. Značen „MX“. Odráží v sobě informace z Mascotu a X!Tandemu zároveň.

Výstupem práce je program realizující systém zpracování skóre ze dvou metod identifikace proteinů v tandemové hmotnostní spektrometrii. Program je napsaný v jazyce prostředí MATLAB verze 7.2.0.232 (R2006a). A realizovaný ve formě funkce pro MATLAB.

LITERATURA

[1] LEXA, Matěj. Obrázek z prezentace – Buňka DNA u člověka. [Online] 2007. < www.fi.muni.cz/~lexa >. Dostupné na WWW:
[2] Červená krvinka. [Online] < www.technet.idnes.cz >.
[3] Wikipedia . [Online] 2008. Dostupné na WWW: < http://en.wikipedia.org >.
[4] Léčba AIDS upravenými imunitními buňkami. *Objective source e-learning.* [Online] 2007. Dostupné na WWW: < http://www.osel.cz >.
[5] KOZUPLÍK, J. *Prezentace - Schéma párování a spojení šroubovice DNA.* 2007.
[6] Schéma - Struktura pro složení chromozomu. *Human Genome Project .* [Online] U. S. Department of Energy Office of Science. Dostupné na WWW: < http://www.ornl.gov/hgmis >.
[7] CHURÝ, Lukáš. Historie a zaměření bioinformatiky, Struktura a funkce DNA, Geny, genomy a buňky. *Bioinformatika.* [Online] 2008. Dostupné na WWW: < http://programujte.com >.
[8] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky.* Praha : ACADEMIA, 2006. ISBN 80-200-1360-1.
[9] ZRZAVÝ. *Sekvenování DNA.* 2004.
[10] Mascot. [Online] Matrixscience, 2008. [Citace: 11. 5 2008.] Dostupné na WWW: < http://www.matrixscience.com >.
[11] VALLA, Martin. *Dynamické programování v analýze genetických dat*. Brno, 2007.
[12] HOLČAPEK, Michal. Hmotnostní spektrometrie v organické analýze. [Online] Mass Spectrometry Group, 2008. Dostupné na WWW: < http://holcapek.upce.cz/teaching_CZ.htm >.
[13] HAVLIŠ, Jan. Hmotnostní spektrometrie MALDI TOF. [Online] Vesmír, 8 1999. Dostupné na WWW: < http://www.vesmir.cz >.
[14] MYŠKOVÁ H., VACKOVÁ, M., CHRASTINA, P., ŠŤASTNÁ S. Využití tandemové hmotnostní spektrometrie. *Abstrakt konference BioLab.* [Online] 2003. Dostupné na WWW: < http://ukb.lf1.cuni.cz/biolab/abs_biol.htm >.
[15] Sekvenování bílkovin a peptidů. *MALDI-TOF Mass Spec Group.* [Online] Vysoká škola chemicko-technologická v Praze, 2005. [Citace: 11. 5 2008.] Dostupné na WWW: < http://biomikro.vscht.cz/maldiman/cz/theory/sequencing.php >.

[16] ŠEDA, Ondřej. LIŠKA, František. ŠEDOVÁ Lucie**.** Úvod do proteomiky. *Multimediální učebnice lékařské biologie, genetiky a genomiky.* [Online] Autorský kolektiv, 2006. [Citace: 11. 5 2008.] Dostupné na WWW: < http://biol.lf1.cuni.cz/ucebnice/autori.htm >.

[17] Overwinter. *Expression Proteomics at Conway Institute UCD.* [Online] Proteomics Group in University College Dublin, 2007. Dostupné na WWW: < http://proteomics.ucd.ie/overwinter >.

[18] RAW DATA AVAILABLE FOR DOWNLOAD. *Peptide Atlas.* [Online] Istitute for Systems Biology, 2008. [Citace: 11. 5 2008.] Dostupné na WWW: < http://www.peptideatlas.org/repository >.

[19] Data repository. *Collection of MS experiments in both the vendor native and the mzXML format.* [Online] Sourceforge.net, 05 11, 2008. [Cited: 05 11, 2008.] Dostupné na WWW: < http://sashimi.sourceforge.net >.

[20] Mass Spectrometry Standards Working Group . *PSI-MS.* [Online] HUPO Proteomics Standards Initiative , 2008. [Citace: 11. 5 2008.] Dostupné na WWW: < http://www.psidev.info >.

[21] PAPPIN, D.J., HOJRUP, P., BLEASBY, A. J. *Rapid identification of proteins by peptide-mass fingerprinting.* 1993. stránky Current Biology 1;3(6):327-32.

[22] KOTYZA, Jaromír. Webové služby pro distribuci číselníků datového standardu. *Datový standard MZ ČR - verze 4.* [Online] MZ ČR, 2008. Dostupné na WWW: < http://ciselniky.dasta.mzcr.cz >.

[23] Mikhail M. Savitski, Michael L. Nielsen and Roman A. Zubarev. New Data Base-independent, Sequence Tag-based Scoring of Peptide MS/MS Data Validates Mowse Scores, Recovers Below Threshold Data, Singles Out Modified Peptides, and Assesses the Quality of MS/MS Techniques. [Online] 2005. Dostupné na WWW: < http://www.mcponline.org >.

[24] Mascot Identity Score. [Online] Proteome software, 2008. [Citace: 11. 5 2008.] Dostupné na WWW: < https://proteome-software.wikispaces.com/Mascot_Identity_Score >.

[25] The Statistics of Sequence Similarity Scores. [Online] NCBI. [Citace: 11. 5 2008.] Dostupné na WWW: < http://www.ncbi.nlm.nih.gov/BLAST/tutorial >.

[26] Trans-Proteomic Pipeline . *Seattle Proteome Center* . [Online] System Biology. [Citace: 11. 5 2008.] Dostupné na WWW: < http://tools.proteomecenter.org >.

[27] *TANDEM project.* [Online] The Global Proteome Machine Organization, 2008. Dostupné na WWW: < http://www.thegpm.org/TANDEM/index.html > .

[28] Craig, R. and Beavis,R.C. *An explanation of the X!Tandem MS/MS spectra search program.* 2003. stránky 2310–2316.

[29] Searle, Brian C. X!Tandem. [Online] Dostupné na WWW: < www.proteomesoftware.com >.

[30] Tandem Mass Spectrometry Protein Identification on a PC Grid. [Online] 2007. Dostupné na WWW:
< http://infoscience.epfl.ch/record/101150/files/ >.
[31] BIOML standard. *BIOML*. [Online] 1999. Dostupné na WWW:
< http://www.proteometrics.com/BIOML > .
[32] Generalized Analytical Merkup Language. *GAML format.* [Online] Thermo Fisher Scientific, 2007. Dostupné na WWW:
< http://www.gaml.org/default.asp >.
[33] Search form . *GPM Tornado.* [Online] The Experimental Bioinformatics Laboratory. [Citace: 11. 5 2008.] Dostupné na WWW:
< http://human.thegpm.org/tandem/thegpm_tandem.html > .
[34] Proteomika. *Proteomická sekce ČSBMB.* [Online] Laboratory of Bioinformatics, Institute of Microbiology, Academy of Sciences of the Czech Republic, Prague, 2001. Dostupné na WWW:
< http://proteom.biomed.cas.cz/proteomics/proteomics.cs.php >.
[35] Proteomika . *Výuka a vzdělávání.* [Online] ÚHKT. [Citace: 11. 5 2008.] Dostupné na WWW:
< http://www.uhkt.cz/vyuka/proteomika >.
[36] Zkratky a odkazy. [Online] SEKK spol. s r.o. [Citace: 11. 05 2008.] Dostupné na WWW:
< http://www.eqa.cz/terminologie/Text/Zkratky.htm >.
[37] *PSpad - editor for developers.* [Online] PSpad, 2008.
< http://www.pspad.com >.
[38] ABZ.cz: slovník cizích slov . [Online] ABC - Radek Kučera, 2006. Dostupné na WWW:
< http://slovnik-cizich-slov.abz.cz >.
[39] EDDY, 2004, popis sestavení BLOSUM.
[40] Bioinformatics explained: Scoring matrices, Science. Dostupné na WWW:
<http://clcbio.com >.
[41] Internetový ochod firmy LabX: hmotnostní spektrometry. [Online] LABX, 2008. Dostupné na WWW:
< http://www.labx.com >.

SEZNAM ZKRATEK

DNA	kyselina deoxyribonukleová
RNA	kyselina ribonukleová
IUPAC	Mezinárodní unie pro čistou a aplikovanou chemii
MS	hmotnostní spektrometrie
MS/MS	tandemová hmotnostní spektrometrie
TOF	doba letu
MX	metaskóre